



US005724352A

# United States Patent [19]

Cloonan et al.

[11] Patent Number: 5,724,352  
 [45] Date of Patent: \*Mar. 3, 1998

[54] **TERABIT PER SECOND PACKET SWITCH HAVING ASSIGNABLE MULTIPLE PACKET LOSS PROBABILITIES**

[75] Inventors: Thomas Jay Cloonan, Downers Grove; Gaylord Warner Richards, Lisle, both of Ill.

[73] Assignee: Lucent Technologies Inc., Murray Hill, N.J.

[\*] Notice: The term of this patent shall not extend beyond the expiration date of Pat. No. 5,544,160.

[21] Appl. No.: 522,209

[22] Filed: Aug. 31, 1995

[51] Int. Cl<sup>6</sup> H04L 12/54

[52] U.S. Cl. 370/395; 370/413; 370/419; 340/825.5

[58] Field of Search 370/252, 253, 370/229, 230, 231, 235, 238, 395, 396, 397, 398, 399, 412, 413-419, 420; 340/825.5, 825.51, 825.52

## [56] References Cited

### U.S. PATENT DOCUMENTS

4,035,584	7/1977	Lurtz	.....	179/15
4,592,048	5/1986	Beckner et al.	.....	370/354
4,686,669	8/1987	Chang	.....	370/375
4,872,158	10/1989	Richards	.....	370/380
4,988,993	1/1991	Hwang et al.	.....	340/825
4,993,016	2/1991	Richards	.....	370/351
5,040,173	8/1991	Richards	.....	370/351
5,077,483	12/1991	Cloonan et al.	.....	359/135
5,122,892	6/1992	Cloonan et al.	.....	359/117
5,258,978	11/1993	Cloonan et al.	.....	370/411
5,291,482	3/1994	McHarg et al.	.....	370/413
5,311,345	5/1994	Cloonan et al.	.....	359/139

5,345,441	9/1994	Paker et al.	.....	370/358
5,390,176	2/1995	Schoute et al.	.....	370/395
5,453,981	9/1995	Katsube et al.	.....	370/397
5,544,160	8/1996	Cloonan et al.	.....	370/395
5,550,823	8/1996	Irie et al.	.....	370/413
5,583,858	12/1996	Hanaoka	.....	370/395

## OTHER PUBLICATIONS

Y. S. Yeh et al., "The Knockout Switch: A Simple, Modular Architecture for High-Performance Packet Switching", *ISS '87 AT&T Technical Papers*, pp. 287-311.

W. B. Stephens et al., "Terabit-per-Second Throughput Switches for Broadband Central Offices: An Overview", *IEEE LCS*, Nov. 1990, pp. 20-26.

K.Y. Eng et al., "A Growable Packet (ATM) Switch Architecture: Design Principles and Applications", *IEEE Transactions on Communications*, vol. 40, No. 2, Feb. 1992, pp. 423-430.

Primary Examiner—Benedict V. Safourek

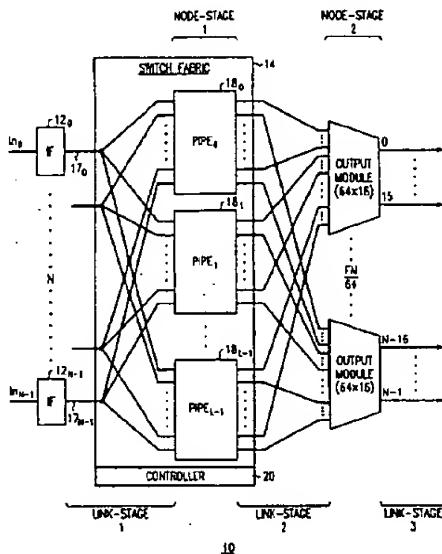
Assistant Examiner—Seema S. Rao

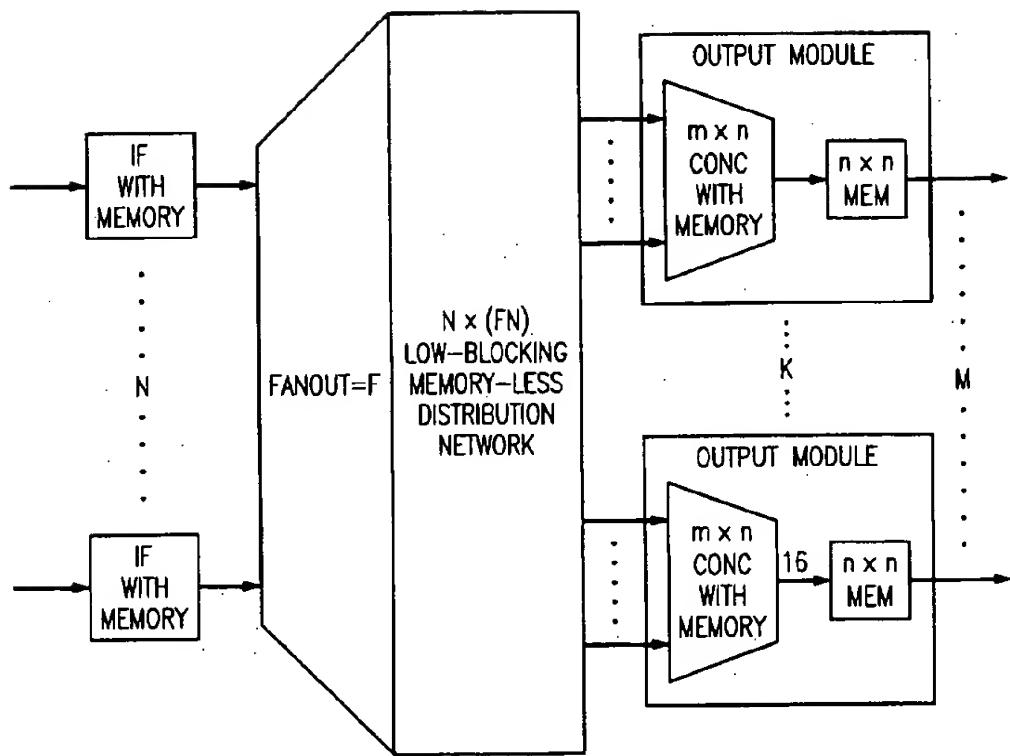
Attorney, Agent, or Firm—Jack R. Penrod

## [57] ABSTRACT

A physically realizable one terabit or more ATM packet switch that has a large number of input interfaces connected to a single stage switching fabric which is in turn connected to a number of output modules, generally according to the growable packet switch architecture. This ATM packet switch is different from other growable packet switches in that it has a single stage switch fabric controlled by an out-of-band controller, yet it has significantly reduced complexity with respect to comparably sized electronic crossbar switches or their isomorphs. The out-of-band controller has multiple priority levels in order to provide high priority users with a near certainty that their packets will be successfully routed, while delivering an acceptably low packet or cell loss probability to users at the lowest priority level.

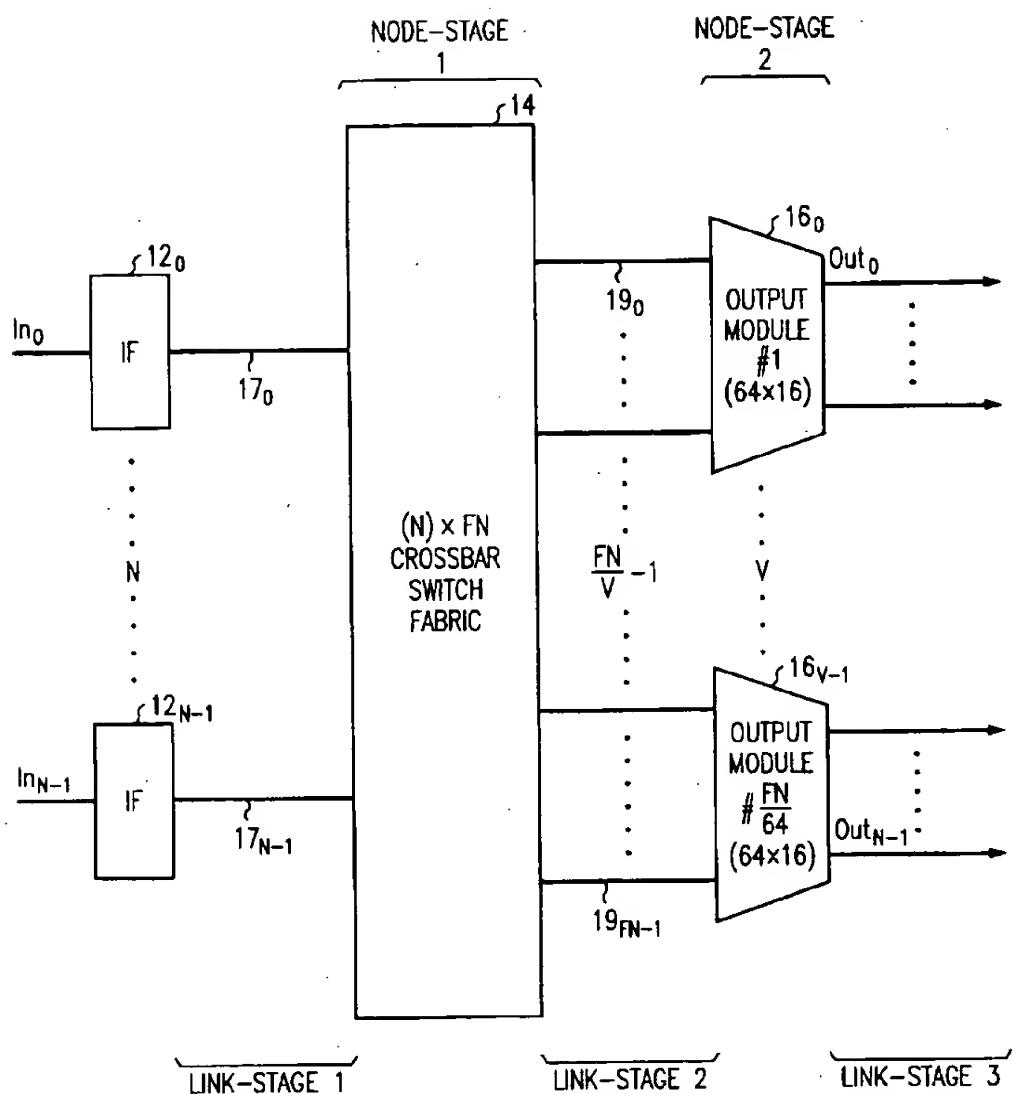
21 Claims, 17 Drawing Sheets





(PRIOR ART)

**FIG. 1**



10

FIG. 2

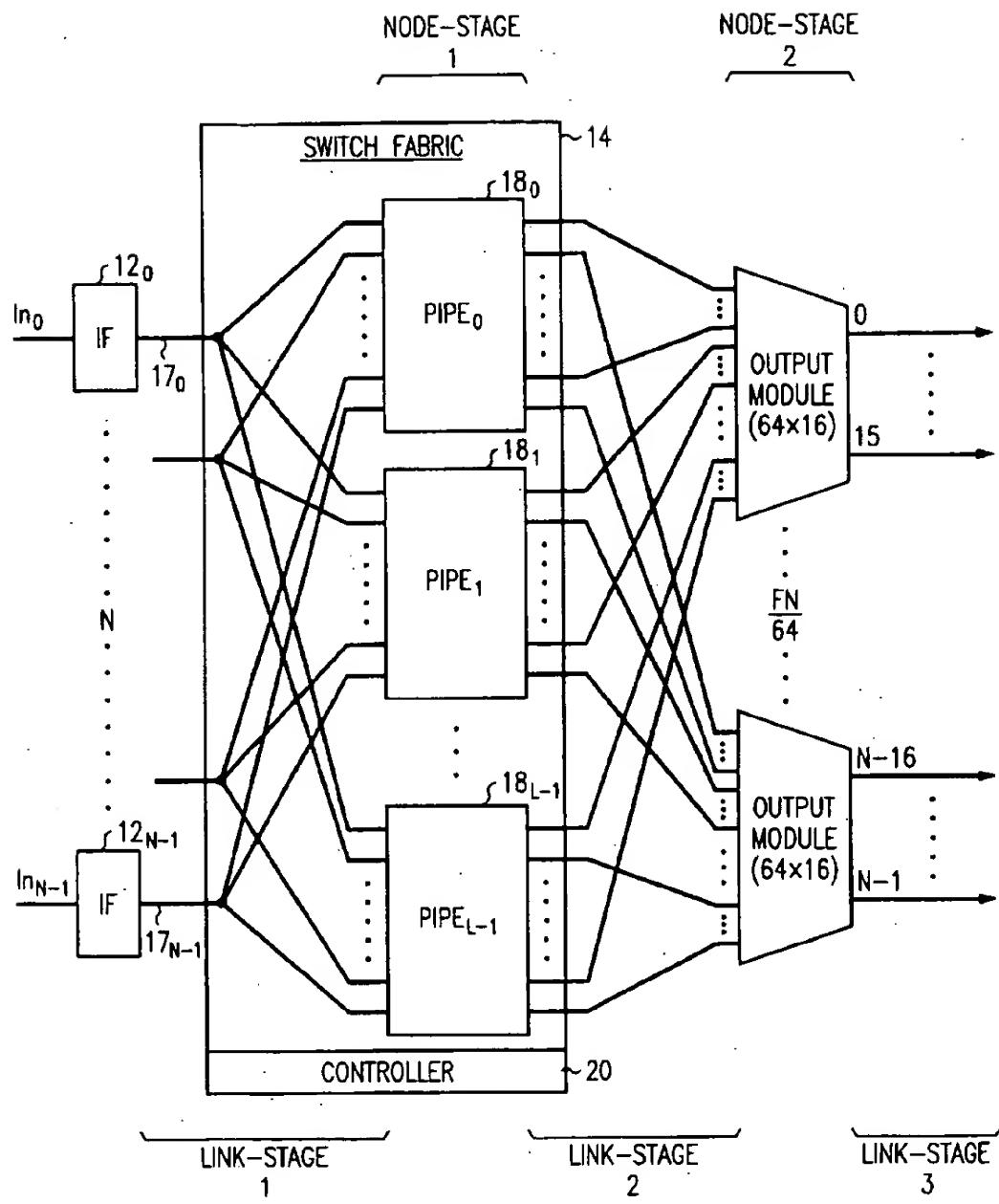


FIG. 3

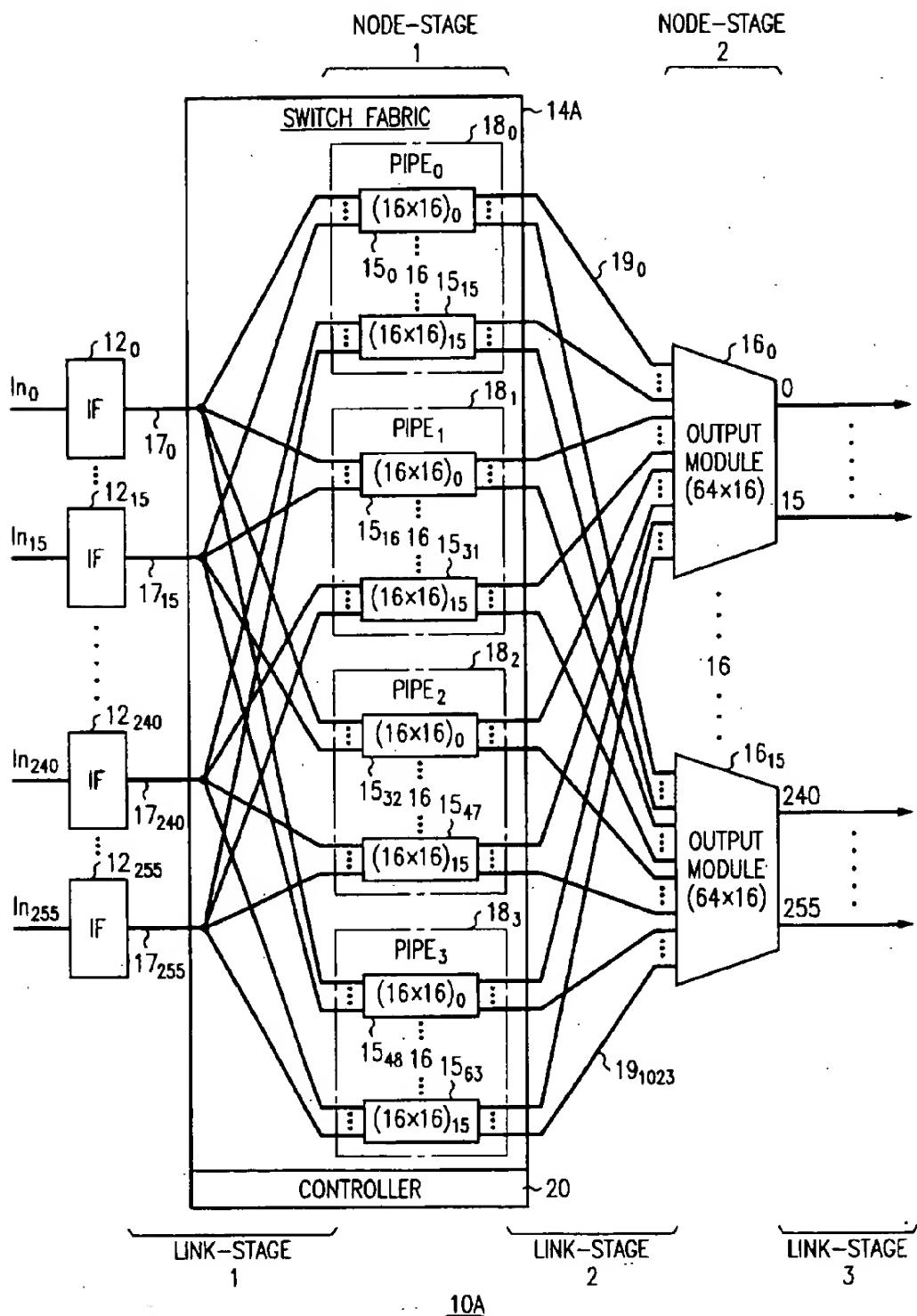
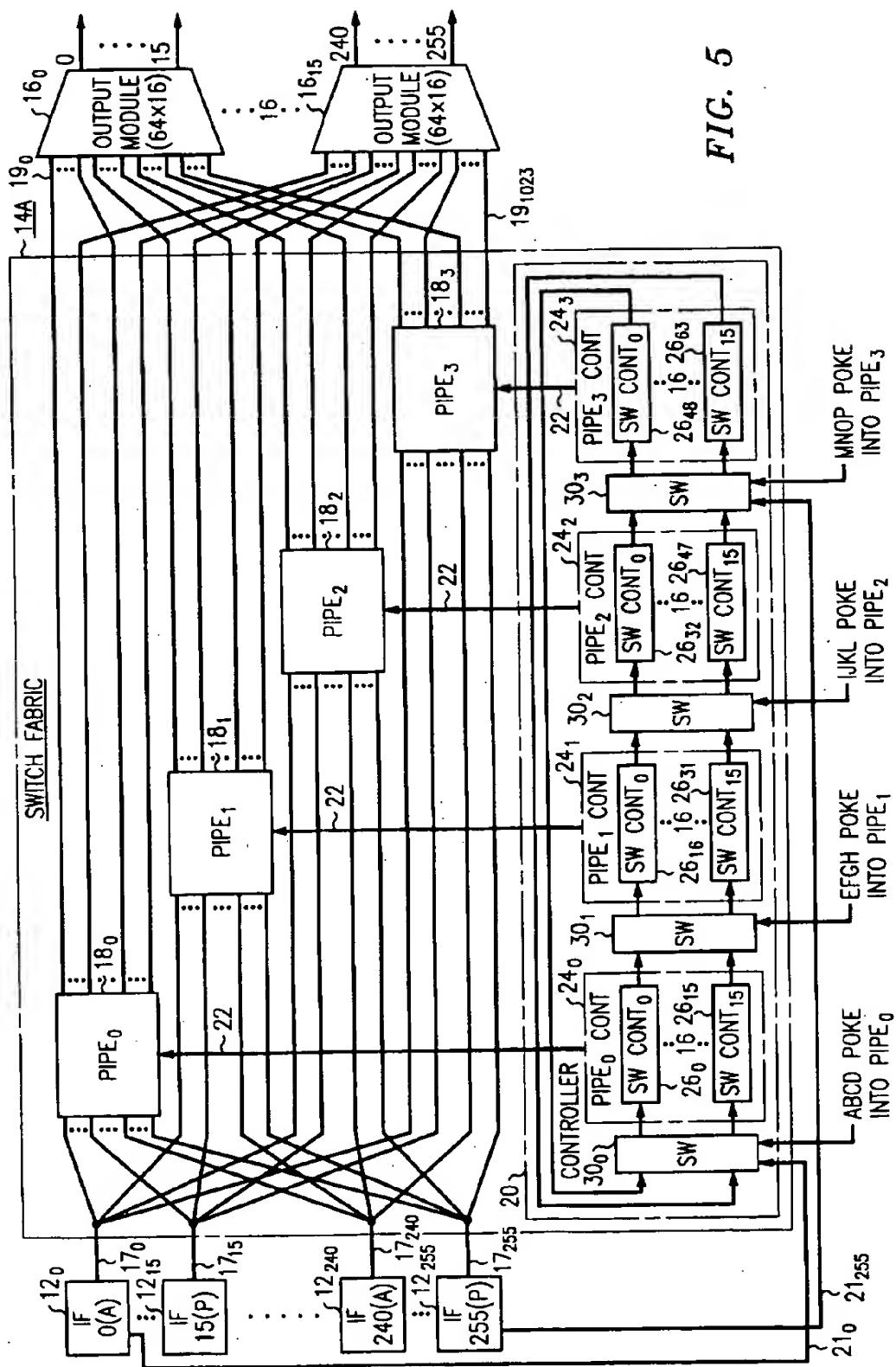


FIG. 4



TIME	PIPE <sub>0</sub> CONTROLLER	PIPE <sub>1</sub> CONTROLLER	PIPE <sub>2</sub> CONTROLLER	PIPE <sub>3</sub> CONTROLLER
i, 0	GROUP 1, PERIOD i			
i, 1		GROUP 1, PERIOD i		
i, 2		GROUP 2, PERIOD i	GROUP 1, PERIOD i	
i, 3			GROUP 2, PERIOD i	GROUP 1, PERIOD i
i, 4			GROUP 3, PERIOD i	GROUP 2, PERIOD i
i, 5	GROUP 2, PERIOD i+1			GROUP 3, PERIOD i
i, 6	GROUP 3, PERIOD i+1			GROUP 4, PERIOD i
i, 7	GROUP 4, PERIOD i+1	GROUP 3, PERIOD i+1		
i, 8		GROUP 4, PERIOD i+1		
i, 9			GROUP 4, PERIOD i+1	

FIG. 6

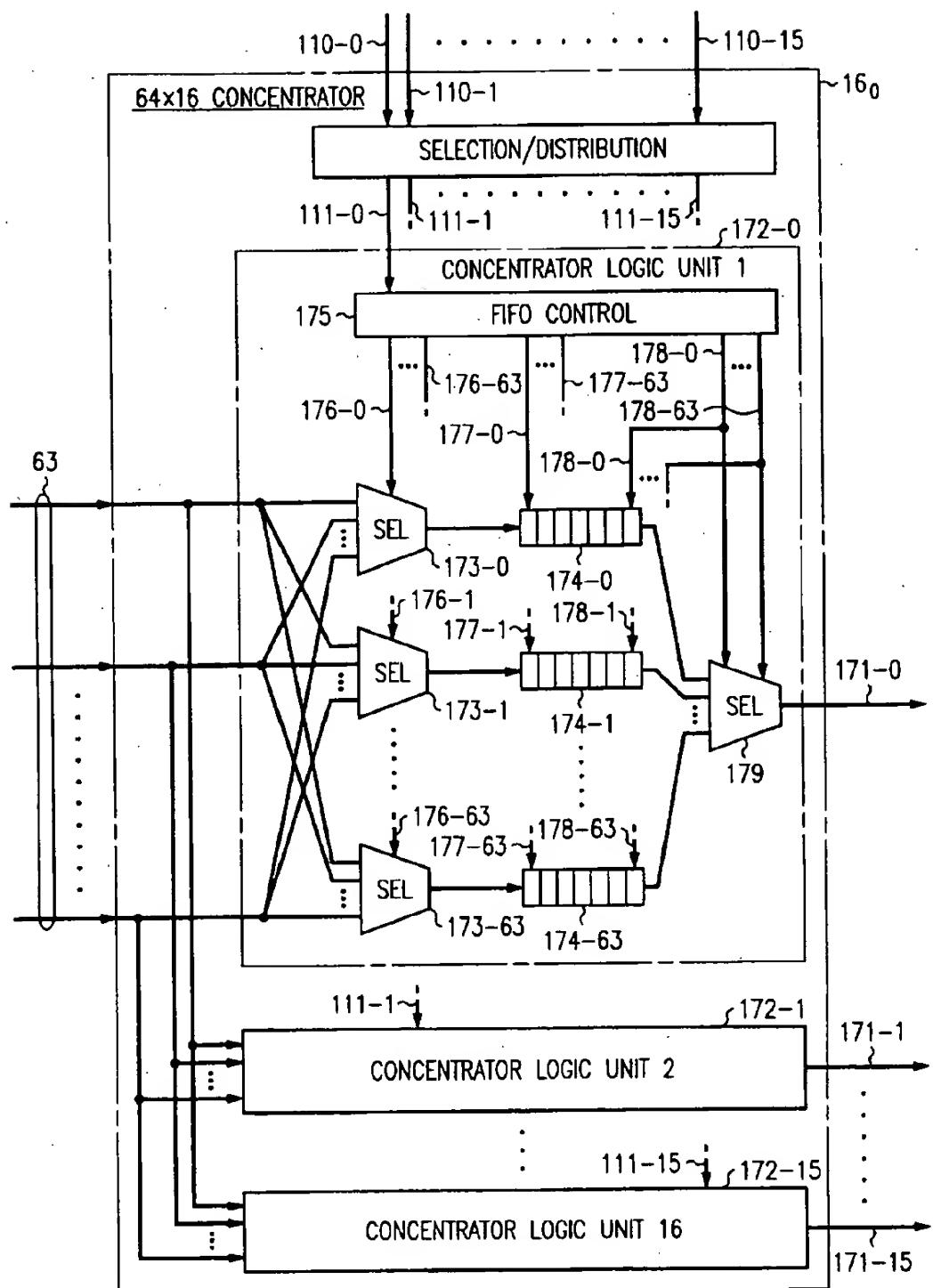


FIG. 7

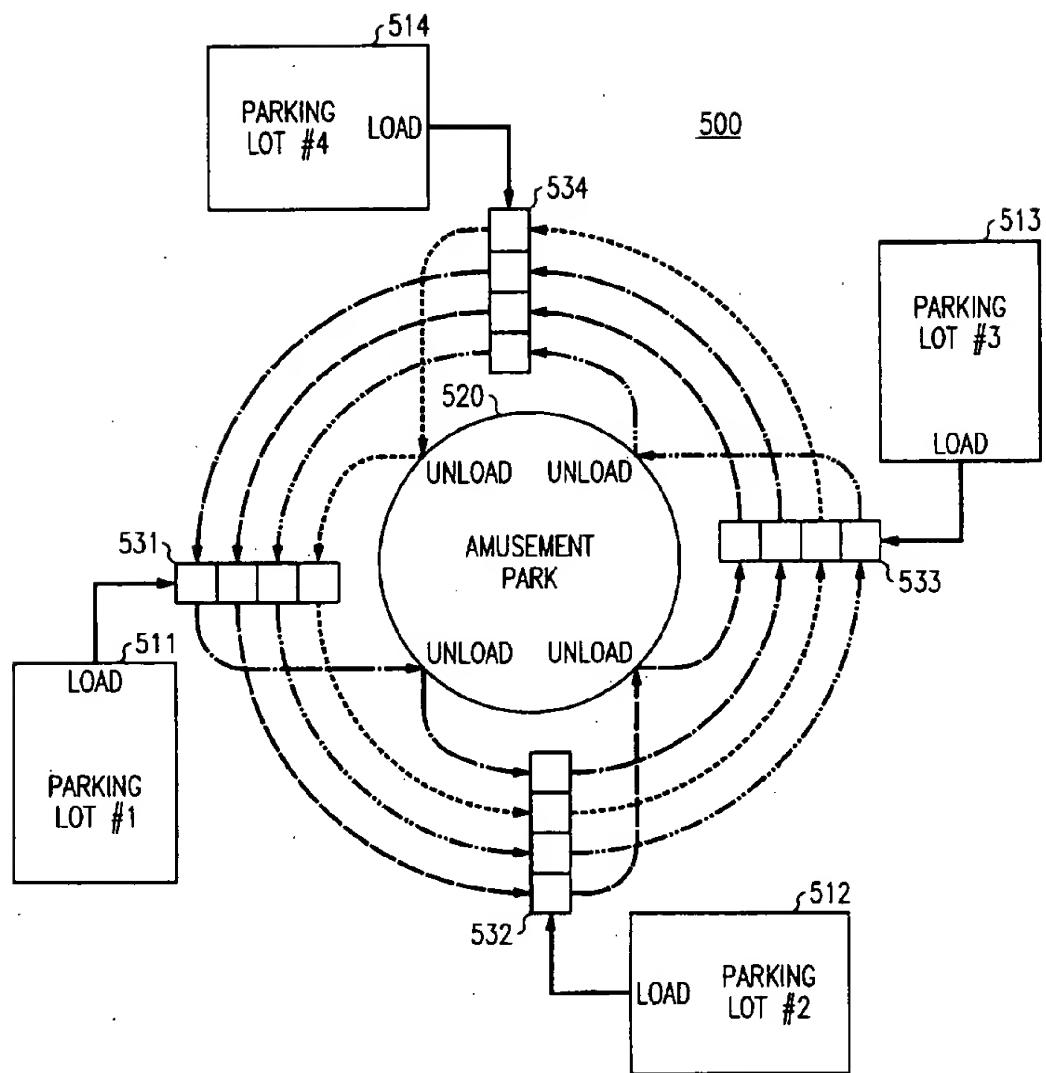


FIG. 8

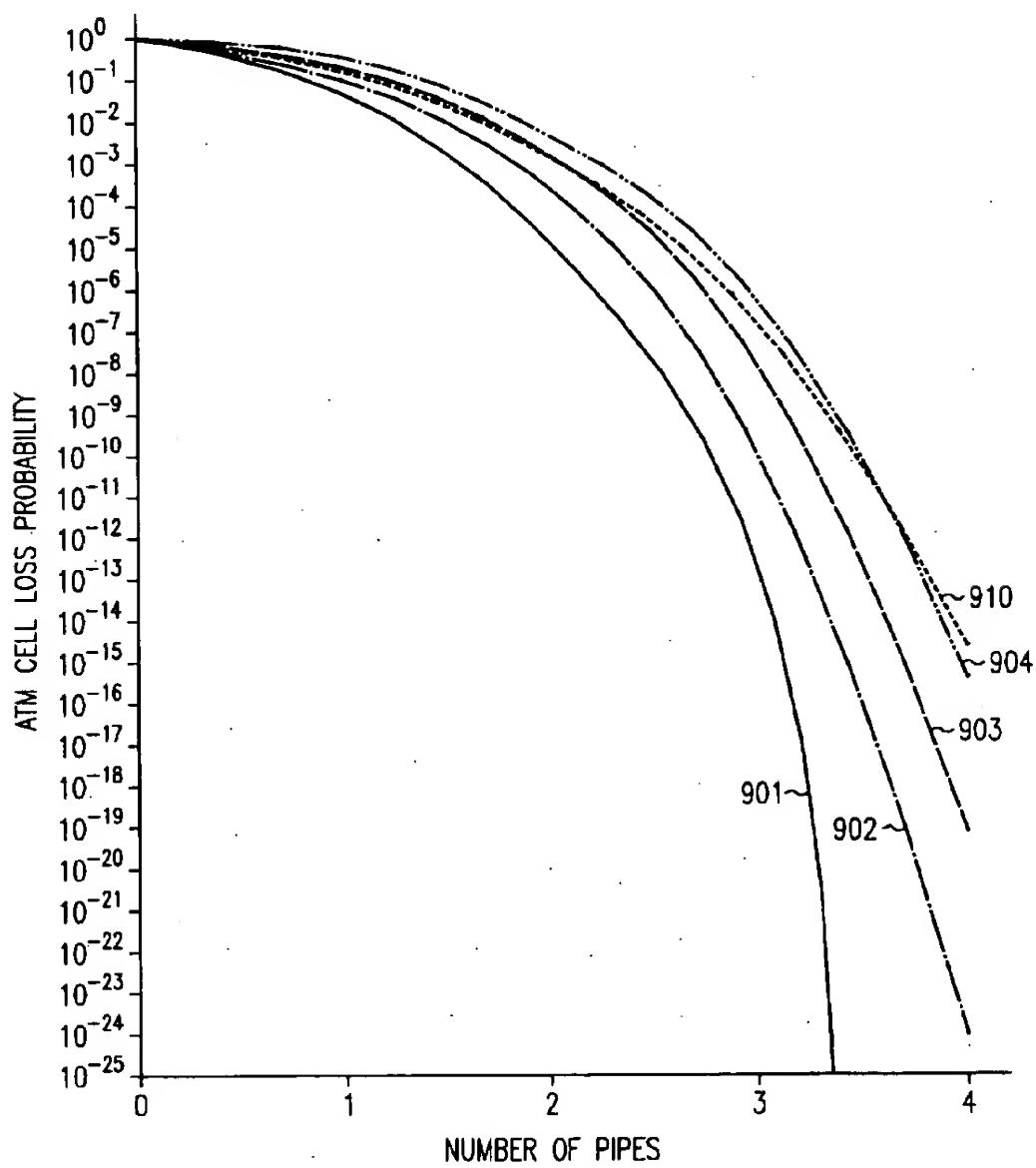


FIG. 9

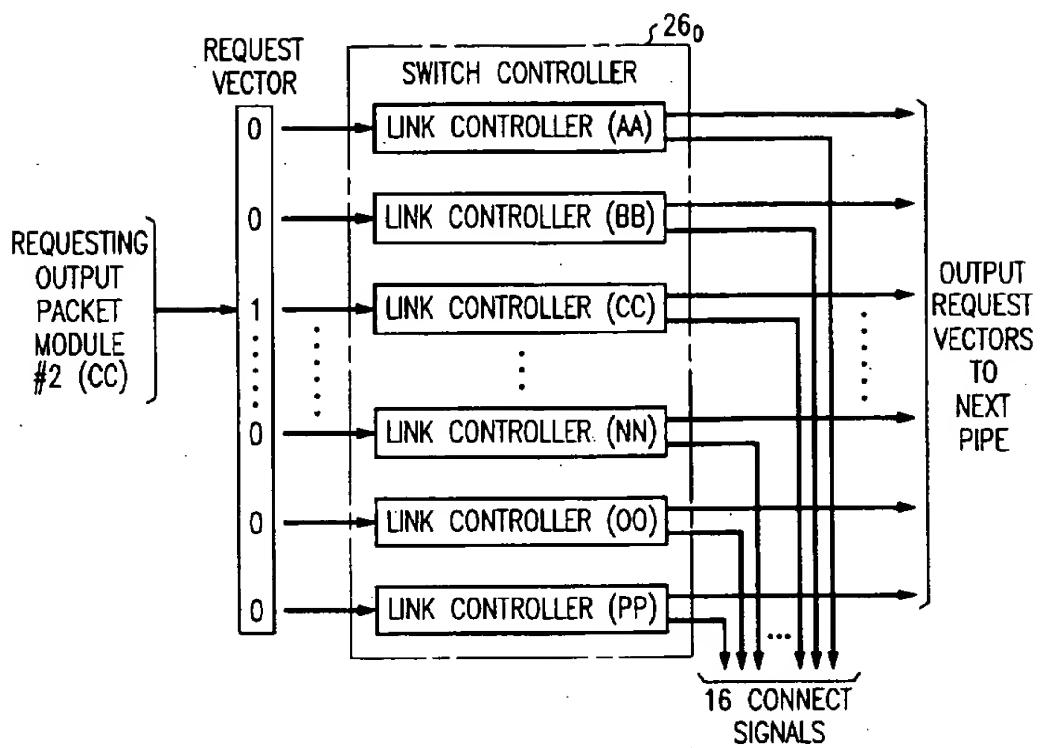


FIG. 10

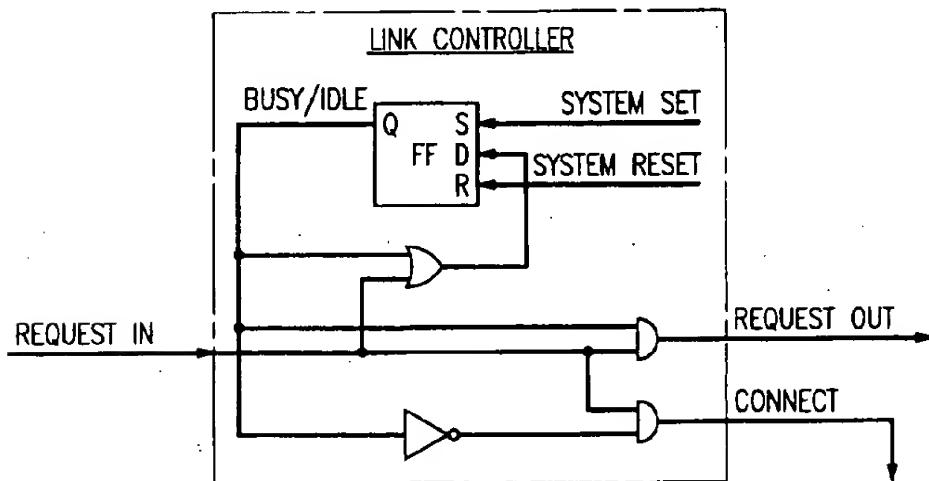


FIG. 11

*FIG. 12*

REQUEST IN	CURRENT BUSY/IDLE	REQUEST OUT	CONNECT	NEXT BUSY/IDLE
0	0	0	0	0
0	1	0	0	1
1	0	0	1	1
1	1	1	0	1

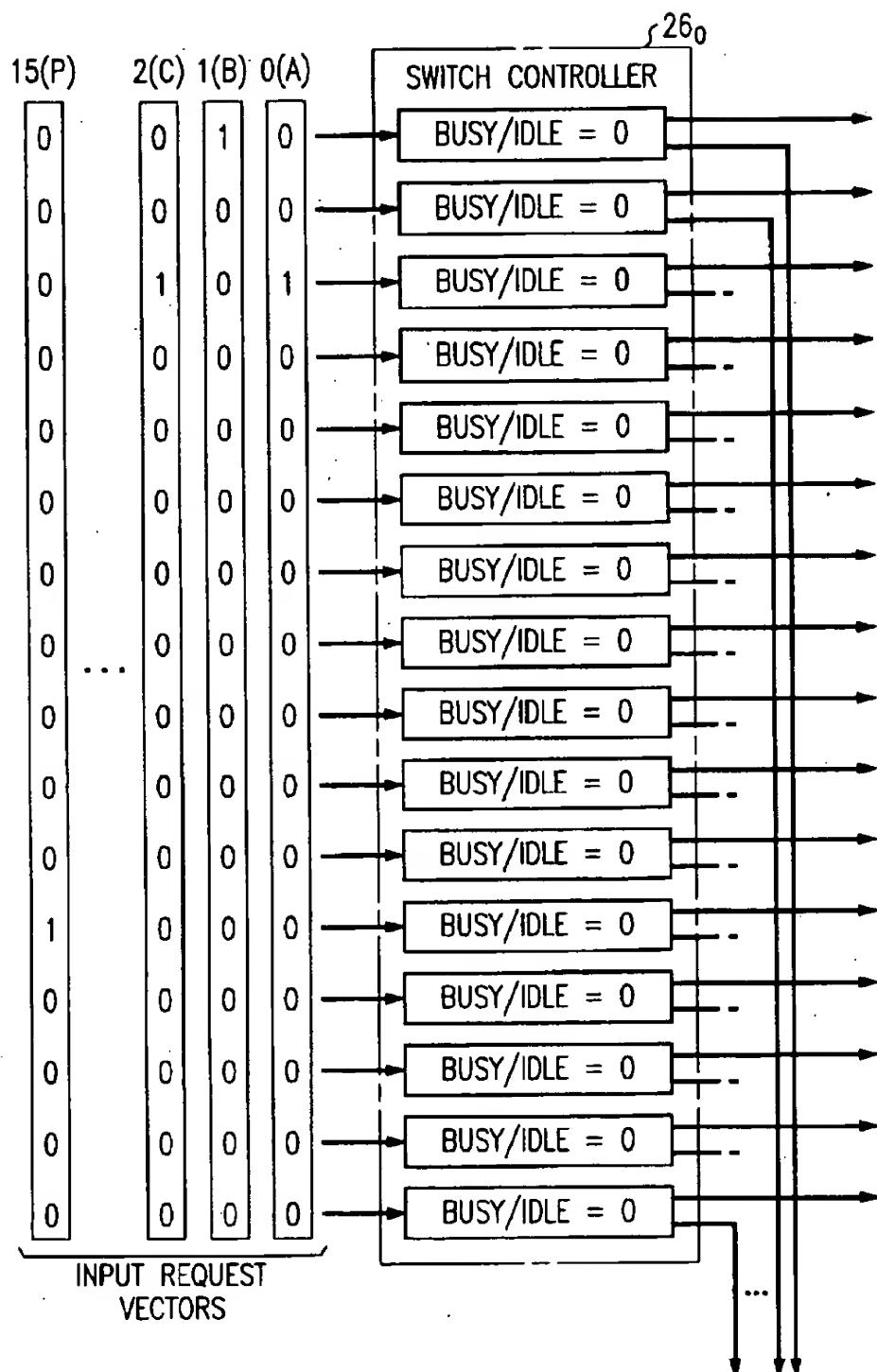


FIG. 13A

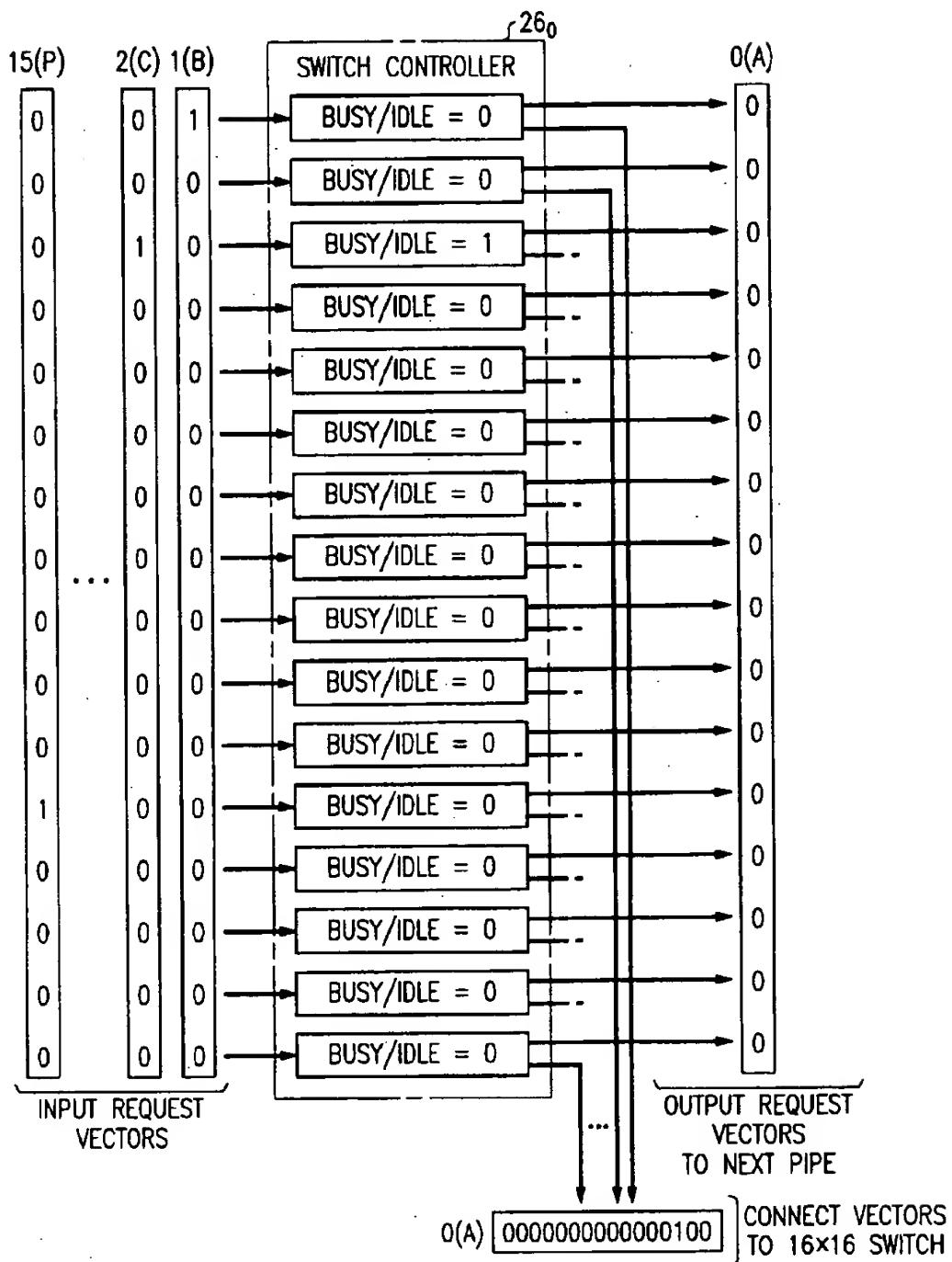


FIG. 13B

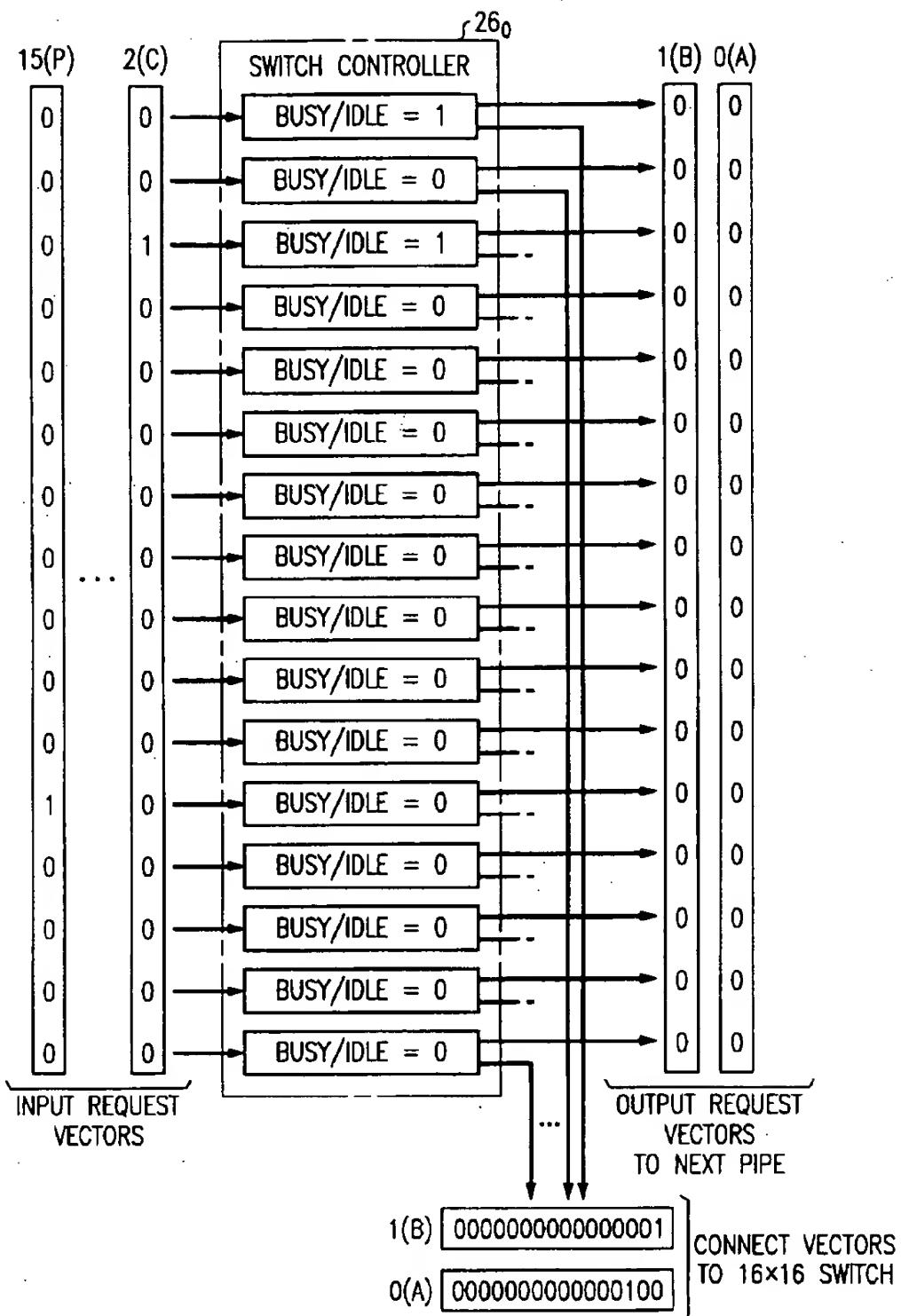


FIG. 13C

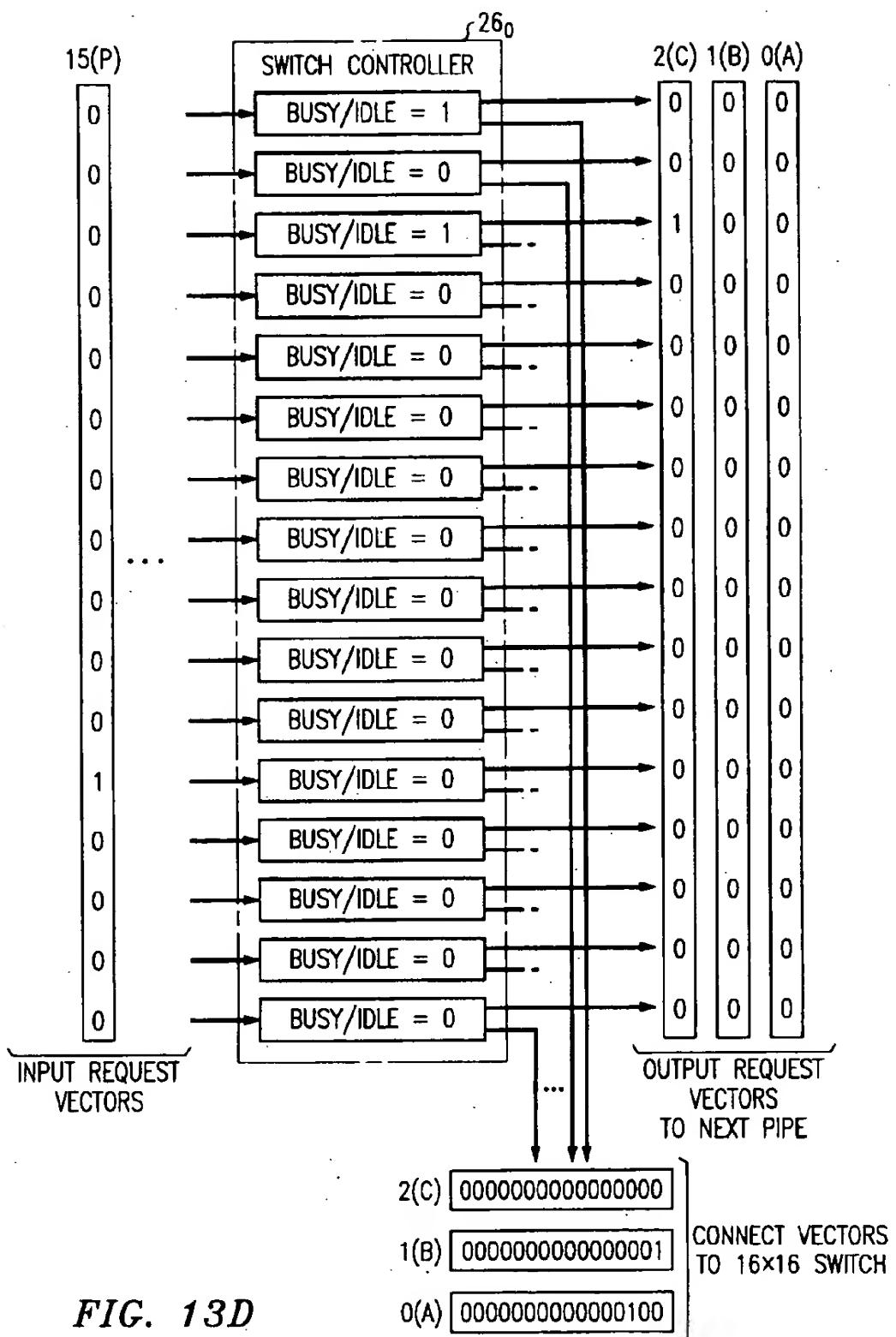


FIG. 13D

FIG. 14

TIME	CONTROLLER 24 <sub>0</sub>	CONTROLLER 24 <sub>1</sub>	CONTROLLER 24 <sub>2</sub>	CONTROLLER 24 <sub>3</sub>
i-1,17	$R_{N,i-1} \times B_{0,i}$	$R_{M,i-1} \times B_{1,i}$	* 2	$R_{O,i-1} \times B_{3,i-1}$
i-1,18	$R_{O,i-1} \times B_{0,i}$	$R_{N,i-1} \times B_{1,i}$	$R_{M,i-1} \times B_{2,i}$	$R_{P,i-1} \times B_{3,i-1}$
i-1,19	$R_{P,i-1} \times B_{0,i}$	$R_{O,i-1} \times B_{1,i}$	$R_{N,i-1} \times B_{2,i}$	
i,0	$R_{A,i} \times B_{0,i}$	$R_{P,i} \times B_{1,i}$	$R_{O,i} \times B_{2,i}$	
i,1	$R_{B,i} \times B_{0,i}$	$R_{A,i} \times B_{1,i}$	$R_{P,i} \times B_{2,i}$	
i,2	$R_{C,i} \times B_{0,i}$	$R_{B,i} \times B_{1,i}$	$R_{A,i} \times B_{2,i}$	* 3
i,3	$R_{D,i} \times B_{0,i}$	$R_{C,i} \times B_{1,i}$	$R_{B,i} \times B_{2,i}$	$R_{A,i} \times B_{3,i}$
i,4		$R_{D,i} \times B_{1,i}$	$R_{C,i} \times B_{2,i}$	$R_{B,i} \times B_{3,i}$
i,5		$R_{E,i} \times B_{1,i}$	$R_{D,i} \times B_{2,i}$	$R_{C,i} \times B_{3,i}$
i,6		$R_{F,i} \times B_{1,i}$	$R_{E,i} \times B_{2,i}$	$R_{D,i} \times B_{3,i}$
i,7	* 0	$R_{G,i} \times B_{1,i}$	$R_{F,i} \times B_{2,i}$	$R_{E,i} \times B_{3,i}$
i,8	$R_{E,i} \times B_{0,i+1}$	$R_{H,i} \times B_{1,i}$	$R_{G,i} \times B_{2,i}$	$R_{F,i} \times B_{3,i}$
i,9	$R_{F,i} \times B_{0,i+1}$		$R_{H,i} \times B_{2,i}$	$R_{G,i} \times B_{3,i}$
i,10	$R_{G,i} \times B_{0,i+1}$		$R_{I,i} \times B_{2,i}$	$R_{H,i} \times B_{3,i}$
i,11	$R_{H,i} \times B_{0,i+1}$		$R_{J,i} \times B_{2,i}$	$R_{I,i} \times B_{3,i}$
i,12	$R_{I,i} \times B_{0,i+1}$	* 1	$R_{K,i} \times B_{2,i}$	$R_{J,i} \times B_{3,i}$
i,13	$R_{J,i} \times B_{0,i+1}$	$R_{I,i} \times B_{1,i+1}$	$R_{L,i} \times B_{2,i}$	$R_{K,i} \times B_{3,i}$
i,14	$R_{K,i} \times B_{0,i+1}$	$R_{J,i} \times B_{1,i+1}$		$R_{L,i} \times B_{3,i}$
i,15	$R_{L,i} \times B_{0,i+1}$	$R_{K,i} \times B_{1,i+1}$		$R_{M,i} \times B_{3,i}$
i,16	$R_{M,i} \times B_{0,i+1}$	$R_{L,i} \times B_{1,i+1}$		$R_{N,i} \times B_{3,i}$
i,17	$R_{N,i} \times B_{0,i+1}$	$R_{M,i} \times B_{1,i+1}$	* 2	$R_{O,i} \times B_{3,i}$
i,18	$R_{O,i} \times B_{0,i+1}$	$R_{N,i} \times B_{1,i+1}$	$R_{M,i} \times B_{2,i+1}$	$R_{P,i} \times B_{3,i}$
i,19	$R_{P,i} \times B_{0,i+1}$	$R_{O,i} \times B_{1,i+1}$	$R_{N,i} \times B_{2,i+1}$	
i+1,0	$R_{A,i+1} \times B_{0,i+1}$	$R_{P,i} \times B_{1,i+1}$	$R_{O,i} \times B_{2,i+1}$	
i+1,1	$R_{B,i+1} \times B_{0,i+1}$	$R_{A,i+1} \times B_{1,i+1}$	$R_{P,i} \times B_{2,i+1}$	
i+1,2	$R_{C,i+1} \times B_{0,i+1}$	$R_{B,i+1} \times B_{1,i+1}$	$R_{A,i+1} \times B_{2,i+1}$	* 3
i+1,3	$R_{D,i+1} \times B_{0,i+1}$	$R_{C,i+1} \times B_{1,i+1}$	$R_{B,i+1} \times B_{2,i+1}$	$R_{A,i+1} \times B_{3,i+1}$
i+1,4		$R_{D,i+1} \times B_{1,i+1}$	$R_{C,i+1} \times B_{2,i+1}$	$R_{B,i+1} \times B_{3,i+1}$

FIG. 15

TIME	CONTROLLER 24 <sub>0</sub>	CONTROLLER 24 <sub>1</sub>	CONTROLLER 24 <sub>2</sub>	CONTROLLER 24 <sub>3</sub>
i-1,9		R' <sub>EFCH,i-1</sub> × B <sub>1,i-1</sub>	R' <sub>ABCD,i</sub> × B <sub>2,i-1</sub>	R' <sub>ABCD,i</sub> × B <sub>3,i-1</sub>
i-1,10		R'' <sub>EFCH,i-1</sub> × B <sub>1,i-1</sub>	R'' <sub>EFCH,i-1</sub> × B <sub>2,i-1</sub>	R'' <sub>ABCD,i</sub> × B <sub>3,i-1</sub>
i-1,11	* 0		R' <sub>EFCH,i-1</sub> × B <sub>2,i-1</sub>	R' <sub>EFCH,i-1</sub> × B <sub>3,i-1</sub>
i,0	R' <sub>EFCH,i-1</sub> × B <sub>0,i</sub>		R'' <sub>JKL,i-1</sub> × B <sub>2,i-1</sub>	R'' <sub>EFCH,i-1</sub> × B <sub>3,i-1</sub>
i,1	R'' <sub>EFCH,i-1</sub> × B <sub>0,i</sub>		R' <sub>JKL,i-1</sub> × B <sub>2,i-1</sub>	R' <sub>JKL,i-1</sub> × B <sub>3,i-1</sub>
i,2	R' <sub>JKL,i-1</sub> × B <sub>0,i</sub>	* 1		R'' <sub>JKL,i-1</sub> × B <sub>3,i-1</sub>
i,3	R'' <sub>JKL,i-1</sub> × B <sub>0,i</sub>	R' <sub>JKL,i-1</sub> × B <sub>1,i</sub>		R' <sub>MNOP,i</sub> × B <sub>3,i-1</sub>
i,4	R' <sub>MNOP,i-1</sub> × B <sub>0,i</sub>	R'' <sub>JKL,i-1</sub> × B <sub>1,i</sub>		R'' <sub>MNOP,i</sub> × B <sub>3,i-1</sub>
i,5	R'' <sub>MNOP,i-1</sub> × B <sub>0,i</sub>	R' <sub>MNOP,i-1</sub> × B <sub>1,i</sub>	* 2	
i,6	R' <sub>ABCD,i</sub> × B <sub>0,i</sub>	R'' <sub>MNOP,i-1</sub> × B <sub>1,i</sub>	R' <sub>MNOP,i-1</sub> × B <sub>2,i</sub>	
i,7	R'' <sub>ABCD,i</sub> × B <sub>0,i</sub>	R' <sub>ABCD,i</sub> × B <sub>1,i</sub>	R'' <sub>MNOP,i-1</sub> × B <sub>2,i</sub>	
i,8		R'' <sub>ABCD,i</sub> × B <sub>1,i</sub>	R' <sub>ABCD,i</sub> × B <sub>2,i</sub>	* 3
i,9		R' <sub>EFCH,i</sub> × B <sub>1,i</sub>	R'' <sub>ABCD,i</sub> × B <sub>2,i</sub>	R' <sub>ABCD,i</sub> × B <sub>3,i</sub>
i,10		R'' <sub>EFCH,i</sub> × B <sub>1,i</sub>	R' <sub>EFCH,i</sub> × B <sub>2,i</sub>	R'' <sub>ABCD,i</sub> × B <sub>3,i</sub>
i,11	* 0		R'' <sub>EFCH,i</sub> × B <sub>2,i</sub>	R' <sub>EFCH,i</sub> × B <sub>3,i</sub>
i+1,0	R' <sub>EFCH,i</sub> × B <sub>0,i+1</sub>		R' <sub>JKL,i</sub> × B <sub>2,i</sub>	R'' <sub>EFCH,i</sub> × B <sub>3,i</sub>
i+1,1	R'' <sub>EFCH,i</sub> × B <sub>0,i+1</sub>		R'' <sub>JKL,i</sub> × B <sub>2,i</sub>	R' <sub>JKL,i</sub> × B <sub>3,i</sub>
i+1,2	R' <sub>JKL,i</sub> × B <sub>0,i+1</sub>	* 1		R'' <sub>JKL,i</sub> × B <sub>3,i</sub>
i+1,3	R'' <sub>JKL,i</sub> × B <sub>0,i+1</sub>	R' <sub>JKL,i</sub> × B <sub>1,i+1</sub>		R' <sub>MNOP,i</sub> × B <sub>3,i</sub>
i+1,4	R' <sub>MNOP,i</sub> × B <sub>0,i+1</sub>	R'' <sub>JKL,i</sub> × B <sub>1,i+1</sub>		R'' <sub>MNOP,i</sub> × B <sub>3,i</sub>
i+1,5	R'' <sub>MNOP,i</sub> × B <sub>0,i+1</sub>	R' <sub>MNOP,i</sub> × B <sub>1,i+1</sub>	* 2	
i+1,6	R' <sub>ABCD,i+1</sub> × B <sub>0,i+1</sub>	R'' <sub>MNOP,i</sub> × B <sub>1,i+1</sub>	R' <sub>MNOP,i</sub> × B <sub>2,i+1</sub>	
i+1,7	R'' <sub>ABCD,i+1</sub> × B <sub>0,i+1</sub>	R' <sub>ABCD,i+1</sub> × B <sub>1,i+1</sub>	R'' <sub>MNOP,i</sub> × B <sub>2,i+1</sub>	
i+1,8		R'' <sub>ABCD,i+1</sub> × B <sub>1,i+1</sub>	R' <sub>ABCD,i+1</sub> × B <sub>2,i+1</sub>	* 3
i+1,9		R' <sub>EFCH,i+1</sub> × B <sub>1,i+1</sub>	R'' <sub>ABCD,i+1</sub> × B <sub>2,i+1</sub>	R' <sub>ABCD,i+1</sub> × B <sub>3,i+1</sub>
i+1,10		R'' <sub>EFCH,i+1</sub> × B <sub>1,i+1</sub>	R' <sub>EFCH,i+1</sub> × B <sub>2,i+1</sub>	R'' <sub>ABCD,i+1</sub> × B <sub>3,i+1</sub>
i+1,11	* 0		R'' <sub>EFCH,i+1</sub> × B <sub>2,i+1</sub>	R' <sub>EFCH,i+1</sub> × B <sub>3,i+1</sub>
i+2,0	R' <sub>EFCH,i+1</sub> × B <sub>0,i+2</sub>		R' <sub>JKL,i+1</sub> × B <sub>2,i+1</sub>	R'' <sub>EFCH,i+1</sub> × B <sub>3,i+1</sub>

**TERABIT PER SECOND PACKET SWITCH  
HAVING ASSIGNABLE MULTIPLE PACKET  
LOSS PROBABILITIES**

**CROSS REFERENCES**

This application is related to the following co-pending applications: "A TERABIT PER SECOND PACKET SWITCH" by Thomas Cloonan and Gaylord Richards, Filing Date Dec. 30, 1994, now U.S. Pat. No. 5,544,160, "TERABIT PER SECOND ATM PACKET SWITCH HAVING DISTRIBUTED OUT-OF-BAND CONTROL" by Thomas Cloonan and Gaylord Richards, Filing Date Dec. 30, 1994 now U.S. Pat. No. 5,642,349, "TERABIT PER SECOND PACKET SWITCH HAVING DISTRIBUTED OUT-OF-BAND CONTROL OF CIRCUIT AND PACKET SWITCHING COMMUNICATIONS" by Thomas Cloonan and Gaylord Richards, Filing Date Dec. 30, 1994, now U.S. Pat. No. 5,537,403, "METHOD AND APPARATUS FOR DETECTING AND PREVENTING THE COMMUNICATION OF BIT ERRORS ON A HIGH PERFORMANCE SERIAL DATA LINK" by Thomas Cloonan, Filing Date Dec. 30, 1994, now U.S. Pat. No. 5,566,193 "TERABIT PER SECOND DISTRIBUTION NETWORK" by Thomas Cloonan and Gaylord Richards, Filing Date Dec. 30, 1994, Ser. No. 08/366,707, currently pending, "APPARATUS AND METHOD FOR REDUCING DATA LOSSES IN A GROWABLE PACKET SWITCH" by Thomas Cloonan and Gaylord Richards Filing Date Dec. 30, 1994, now U.S. Pat. No. 5,550,815, and "TERABIT PER SECOND ATM PACKET SWITCH HAVING OUT-OF-BAND CONTROL WITH MULTICASTING" by Thomas Cloonan and Gaylord Richards, Filing Date Aug. 31, 1995; Ser. No. 08/521,676 currently pending.

**TECHNICAL FIELD**

The invention relates to large telecommunication switches that use data packets in order to communicate at aggregate throughputs at the one terabit per second level, and more particularly to such large telecommunication switches that provide a very low cell loss probability to higher priority users and an acceptably low cell loss probability to lower priority users.

**DESCRIPTION OF THE PRIOR ART**

Telecommunications have long used digital switching to encode, multiplex, transmit and decode audio frequencies in order to carry the millions of telephone voice calls of the world. Telecommunication switches for voice calls have grown to very large sizes to keep pace with demand. Most of the switching systems that route and control voice call traffic are called circuit switches, which means that for each call a type of bi-directional circuit, usually audio is set up between the calling party and the called party. The circuit that is set up has the bandwidth and transport timing necessary to simulate a face-to-face conversation without distortion or time delays that are objectionable to the parties.

An alternative to circuit switching is called packet switching. For packet switching, the calling party is responsible for converting the information into one or more packets. This information could be encoded voice, it could be encoded computer data, or it could be encoded video. A number to guide the packet to its destination, i.e. the called party, is typically included in a packet header. The packet switching network then has the task of routing each packet to its respective destination without undue distortion or delay. The called party usually has the equipment to receive the packets and decode the information back into an appropriate form.

The extremely rapid growth of packet switching traffic carrying voice, computer (LAN/WAN), facsimile, image and video data to an ever widening variety of locations, along with the proposals used as part of the National Information Infrastructure, has challenged both the packet switch system architectures and protocols.

Many vendors and service providers have joined forces to define a global standard that permits packet switching services to be provided in a ubiquitous fashion. The result of this coordinated effort has been the rapid development and deployment of an Asynchronous Transfer Mode (ATM) as a standard means of efficiently routing and transporting data packets that have stochastically-distributed arrival rates according to the recent ATM standard. ATM is thus a packet-oriented standard, but unlike most of its data packet predecessors (X.25, frame relay, etc.), ATM uses short, fixed-length, 53-byte packets that are called cells. ATM also uses a very streamlined form of error recovery and flow control relative to its predecessors. In fact, the ATM standard essentially eliminates most error protection and flow control at the link level, leaving these functions to higher level protocols at the edges of the network. This approach permits rapid routing of the short cells with minimal network delay and jitter, making ATM compatible with voice, data and video services. ATM has been embraced by the computer, LAN, and WAN industries, so a seamless packet communication from the source computer through LANs, WANs, and the public-switched network is a reality.

As this level of connectivity becomes available to the entire spectrum of users, the demand and volume of ATM traffic that may be generated in the future is virtually limitless. As a result, the number and size of the switches and cross-connects required to route the projected ATM packet traffic may also grow by phenomenal rates within the next decade. ATM switches and cross-connects for toll and gateway applications may require aggregate bandwidths ranging from 155 gigabits per second (1000 inputs at SONET OC-3 155 Mbps rates) to 2.4 terabits per second (1000 inputs at SONET OC-48 2.4 gigabits per second rates). Additionally, if demand for broadband services to the home and LAN/WAN connectivity through the public-switched network grows as some experts believe, then local telephone exchange carriers may install ATM switches and cross-connects for metropolitan area network (MAN) applications having aggregate bandwidths ranging from 100 Gigabits per second to 775 Gigabits per second.

By necessity, most of the current architectural research and hardware/software development for ATM switches has concentrated on switches with much smaller aggregate bandwidths which meet the more near-term needs of the marketplace. For example, most projections within the LAN/WAN community require aggregate bandwidths ranging from 150 Mbps to 12 Gigabits per second, and most of the published proposals within the telecommunications industry call for aggregate bandwidths ranging from 20 Gigabits per second to 160 Gigabits per second. Most of these ATM packet switching architectures do not scale up to larger sizes. Scaling up present ATM switching architectures produce systems that are cost prohibitive, size prohibitive, and/or physically unrealizable because of limits of the underlying technology.

For example, very common designs for large, high-throughput switches use a multi-stage interconnection network containing multiple stages of switching nodes (node-stages) interconnected by stages of links (link-stages) to provide multiple paths between input ports and output ports. Clos, Banyan and Benes networks are examples of such

networks. A multiple stage network design can yield networks with very high levels of performance (low blocking probabilities, low delay, high degrees of fault tolerance, etc.), and may result in low system-level costs, because network resources (nodes and links) are time-shared by the many different paths that can be set up within the network. Physically realizing a multistage network for a large throughput ATM packet switch is, however, a problem because of routing time and processing power.

Any large, high-throughput ATM switching architecture must address two fundamental issues that profoundly effect the overall performance of the resulting ATM switch. The first of these issues is cell loss due to blocking within the internal links of the distribution network (also known as the switching fabric), and the second is cell loss due to contention for output ports by two or more ATM cells that pass through the switch at the same moment in time. The first issue can usually be solved by designing a network with sufficient switching fabric (nodes and links) such that multiple paths exist between input ports and output ports. As a result, if two or more ATM cells attempt to use the same shared resource (nodes or links) within the switching fabric, each of the cells can usually find a path, thereby eliminating most of the internal network blocking problem. The second issue requires some type of method and apparatus for handling cells simultaneously arriving at the same output port.

A general technique for a switch to handle cells destined for the same output port is analyzed in an article, A Growable Packet Switch Architecture, IEEE Transactions on Communications, February, 1992, by Eng et al. and in another article The Knockout Switch, ISS AT&T Technical Papers, 1987, by Yeh et al. This general design technique, as shown in FIG. 1, segments a packet switch into two distinct parts. An  $N \times (FN)$  distribution network (which provides for  $N$  input ports) and a bank of  $K \times n$  output packet modules (which provide for a total of  $M=Kn$  output ports). Given that each of the links emanating from the distribution network is required to be terminated at one of the inputs to one of the output packet modules, it can be seen that the equation  $FN=Kn$  must be satisfied. In the switch shown in FIG. 1, the switching fabric is a memory-less  $N \times (FN)$  fanout switch whose function is to route an arriving ATM cell to any of the  $m$  inputs on the output packet module connected to the cell's desired output port. The output packet module is a  $m \times n$  switch with buffers that are available for storing cells that must be delayed when two or more cells contend for a particular output port. If the arriving traffic is uniformly distributed across all output ports and if the buffers within the output packet modules are sufficiently large, then the ratio  $m:n$  can always be chosen large enough to force the cell loss probability within the network to be below any desired cell loss probability level. In fact, if the network size ( $N$ ) is large and if  $R$  represents the switch loading, then the cell loss probability of a network with  $m \times n$  output packet modules as shown by Eng et al is given by:

$$P(\text{cell loss}) =$$

$$[1 - m/(nR)] \left[ 1 - \sum_{k=0}^m \{(nR)^k e^{-(nR)}\} / k! \right] + (nR)^m e^{-(nR)} m!.$$

Existing small packet switches have acceptable cell loss probabilities of approximately  $10^{-12}$ , so any loss probability smaller than existing units are considered acceptable. Although there are some users who would be willing to pay for an even smaller ATM cell loss probability if such were available.

Besides the ATM cell losses because of internal contentions, in an ATM packet switch where all of the  $N$  cells arrive simultaneously at the inputs of the distribution network, the cells must be processed at each stage and path hunts must be processed before the next group of  $N$  cells arrives at the network input ports. If for example, the incoming transmission lines support SONET OC-48 2.5 Gigabits per second bit-rates, then the group of  $N$  ATM cells that arrive together must be processed and sent on to the next stage of the pipeline every 176 nano seconds (the duration of an ATM cell on a 2.5 Gigabits per second link). For large values of  $N$ , a substantial amount of processing power is therefore required to complete the path hunt operations for all  $N$  cells. (For an  $N=256$ , then at least  $1.45 \times 10^9$  path hunts must be completed every second, which corresponds to an average processing rate of one path hunt every 684 pico seconds). Present commercial microprocessors can process approximately 100 million instructions per second. If each path hunt took only one instruction, these  $1.45 \times 10^9$  path hunts would require the total processing power of at least 15 of microprocessors. Thus, a path hunt controller based on something other than a single microprocessor will be necessary for a large ATM packet switch.

Two approaches to solving the path hunting problem can be envisioned. One approach uses in-band, i.e. self-routing, control techniques to perform the required path hunts. For in-band control techniques, the connection requests are prepended to the ATM data cells and routed through the switch along the same paths used by the following ATM payload. This approach typically requires processing elements to be distributed throughout all of the nodes in the network. This results in relatively complicated hardware within each node of the network in order to perform localized path hunting operations (on only the cells that pass through that node) when determining how to route the arriving connection requests and ATM cells. The second approach uses out-of-band control techniques whereby the controller and switch fabric are logically separated with this architecture. Connection requests are routed to the path hunting controller before the control signals resulting from the path hunt are injected into the switch fabric to set the paths. This second approach requires that the out-of-band controller have tremendous processing power, (as mentioned above), because of the many path hunt operations that must be performed in a very short period of time.

Since the path hunt operations in switches that use in-band control techniques are only based on localized traffic information and not on global information with respect to all of the switch traffic, the connections resulting from in-band path hunting may not always be routed in optimal fashion. As a result, systems using in-band control techniques often require more switch fabric (stages and nodes) to provide the same operating characteristics as a less expensive switch based on out-of-band control techniques. In addition, out-of-band control ATM switch architectures share many similarities with the partitioning of many existing telecommunication switching and cross-connect products that have centralized controllers, so the development of a system based on a function architecture should yield fewer design problems than an architecture based on newer architectural approach. Thus, an out-of-band control ATM switch should benefit from lower overall hardware costs and have more standard architectural partitioning. On the other hand, the difficulties associated with performing path hunts in an out-of-band controller for  $N$  arriving ATM cells and the time required by a standard partitioned out-of-band controller to perform  $N$  path hunts tends to argue in favor of a newer in-band control switch architecture. For example assuming

a single path hunt requires at least one read from a busy-idle memory and one write to a busy-idle memory,  $N$  path hunts require  $2N$  accesses to memory. If  $N=256$ , then the controller must make 512 memory accesses every 176 nano seconds, so the average memory access time must be 340 pico seconds. Since 340 pico seconds memories are not commercially available, a path hunt scheme different than the present standard architectural portioning is required for any out-of-band controller.

The high probability that large ATM switches will be required coupled with the uncertainties and shortcomings of present ATM architectures demonstrate a strong need in the art for a packet switch architecture that will operate with throughputs at the terabit per second levels and yet may be built using components manufactured by existing technologies. Also, the desire of some users for even lower ATM cell loss probabilities shows a need in the art to provide a large high throughput ATM switch that has multiple ATM cell loss priorities.

It is an object of the present invention to provide an ATM packet switch architecture that has a large aggregate bandwidth.

It is another object of the invention to provide a high performance ATM packet switch that has a relatively low system-level cost.

It is another object of the invention to provide an ATM packet switch with a low cell loss probability with random traffic, bursty traffic, and focused loads that is better than present packet switches.

It is another object of the invention to provide an ATM packet switch with multiple cell loss probabilities, but even the lowest cell loss probability is as good or better than those of known ATM switches.

#### SUMMARY OF THE INVENTION

Briefly stated, in accordance with one aspect of the invention, the foregoing objects are achieved by providing an ATM packet switch architecture having line cards with one ATM cell period storage, a single stage, reduced interconnects switch fabric, an out-of-band controller operating with multiple cell loss priorities, and output packet modules.

In accordance with another aspect of the invention, the foregoing objects are achieved by providing a packet switch for a switching telecommunication packet from any of a number of input lines to any of a number of output lines. This packet switch includes a number of input interfaces, each having an input port connected to a respective input line of the number of input lines and an output port. Each of the output ports is fanned out to  $F$  input ports, where  $F$  of a switching network, which has  $I$  input ports and  $P$  output ports.  $F$  is the fan-out number, which is an integer greater than 1.  $I$  is a number that is an integer multiple of the number of input lines, and  $P$  is a number that is an integer multiple of the number  $I$  of the input ports. The switching network is partitioned into a number,  $C$ , of pipes, where  $C$  is an integer having a value equal to  $P$  divided by  $I$ . A number output modules, which together have a number of output module inputs, are connected to  $P$  output ports of the switching network. Each of the output module inputs is connected to a respective output port of the  $P$  output ports of the switching network. The output modules together have a number plurality of outputs, each of these output module outputs is connected to a respective output line of said number of output lines. Each pipe of the  $C$  pipes has a path from each of the number of input lines that is connectable to a respective output line of the plurality of output lines. Also, the packet switch includes path hunter for hunting a path for

a telecommunication packet from its respective input interface to its desired output line. This path hunter hunts paths for telecommunication packets having at least two different levels of ATM cell loss priorities.

#### BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a block diagram of a generalized growable packet switch;

FIG. 2 is a slightly re-drawn FIG. 1.

FIG. 3 is a block diagram of a growable packet switch in which the switch fabric is partitioned into  $L$  multiple pipes according to the present invention.

FIG. 4 is a block diagram, similar to FIG. 3, of a specific embodiment of the present invention having four pipes ( $L=4$ ) and showing a configuration for the pipes.

FIG. 5 is a simplified block diagram of the embodiment shown in FIG. 4 which shows greater details of the controller.

FIG. 6 illustrates the timing sequences of requests to the controller shown in FIG. 5.

FIG. 7 is a simplified block diagram of an embodiment of an output module.

FIG. 8 is an illustrative example of rolling and its operation in a plan view of an amusement park and its satellite parking lots.

FIG. 9 shows plots of calculated values of various ATM cell loss probabilities both with and without the assignment of preferences.

FIG. 10 is a simplified block diagram of a representative switch controller and its link controllers.

FIG. 11 is a detailed logic diagram of a link controller.

FIG. 12 is a state table for the link controller shown in FIG. 11.

FIGS. 13A-13D when joined together show the operation of a switch controller in response to a sequence of requests.

FIG. 14 illustrates the rolling of path hunting requests through a switch having four pipe controllers with a single level of cell loss priority.

FIG. 15 illustrates the rolling of path hunting requests through a switch having four pipe controllers with two levels of cell loss priorities.

#### DETAILED DESCRIPTION

Referring now to FIG. 2, a large, generalized switch 10 for ATM communications, is shown in block diagram form. ATM switch 10 has a number of input interfaces  $12_0-12_{N-1}$ , a switch fabric 14, and buffered output modules  $16_0-16_{V-1}$ . For ATM operation, input interfaces  $12_0-12_{N-1}$  are high speed digital amplifiers that serve as a matching networks and power amplifiers for fanning out information received on their inputs to multiple input ports of the switch fabric 14. Each of the input interfaces  $12_0-12_{N-1}$  also needs a capability to store one ATM cell, as will be explained below. Similarly for ATM operation, buffered output modules  $16_0-16_{V-1}$  are concentrators that are buffered to reduce packet loss when two or more packets are directed to and contend for the same output of outputs  $Out_0-Out_{N-1}$ .

Switch fabric 14 includes a fanout  $F$  where each of the outputs from the input interfaces  $12_0-12_{N-1}$  is fanned out to  $F$  inputs within switch fabric 14, such that if ATM switch 10 is an  $N \times N$  switch then switch fabric 14 will have  $FN$  internal inputs and  $FN$  outputs to output modules  $16_0-16_{V-1}$ . Output Modules  $16_0-16_{V-1}$  have a fanin or concentration factor of

F in order to convert the FN outputs of the switch fabric 14 to N output module outputs Out<sub>0</sub>-Out<sub>N-1</sub>. Each output module 16<sub>0</sub>-16<sub>N-1</sub> stores arriving ATM packets in FIFO queues, and then routes the ATM packets at the front of each of these FIFO queues to their desired outputs Out<sub>0</sub>-Out<sub>N-1</sub> when the output ports are available.

Switch fabric 14 is a general distribution network which may be a network of switches, specifically crossbar switches, to provide multiple paths from each of its input ports 17<sub>0</sub>-17<sub>N-1</sub> to each of its output ports 19<sub>0</sub>-19<sub>FN-1</sub>. However, it becomes highly impractical to make an N×N switch out of a single crossbar to operate as the switching component of switch fabric 14 when the size of N exceeds 32. Thus, some other way is needed to realize the general architecture shown in FIG. 2.

Referring now to FIG. 3, an ATM switch 10A that is both practical and possible for N inputs where the size of N is at least 256, is shown. Multiple paths from each input 17<sub>0</sub>-17<sub>N-1</sub> through the switch fabric 14A are provided to prevent blocking. These multiple paths are partitioned into groups called pipes with each pipe providing exactly one path between each input port 17<sub>0</sub>-17<sub>N-1</sub> and each output port 19<sub>0</sub>-19<sub>FN-1</sub> within the network. Thus, switch fabric 14A is made up of multiple pipes 18<sub>0</sub>-18<sub>L-1</sub>. The output modules 16<sub>0</sub>-16<sub>N-1</sub> are essentially the same as the output modules shown in FIG. 2.

Switch fabric as seen in co-pending and commonly assigned application entitled TERABIT PER SECOND DISTRIBUTION NETWORK, which is hereby incorporated by reference, is a single stage, memoryless, and non-self routing network. Since the switch fabric 14A is not unconditionally non-blocking as a full N×N crossbar switch would be, a controller 20 is included to hunt for a path through the four pipes for each ATM cell. Since each of the pipes 18<sub>0</sub>-18<sub>L</sub> contains a path that could transport the ATM cell, the real purpose of the controller 20 is to find a path that is not blocked.

For ATM switch 10A, if the number of input lines, N is equal to 256 and if each input line is operated at a standard 2.5 Gigabits per second data rate, its aggregate throughput will be 0.640 terabits per second. Scaling or growing such an ATM switch by a factor of two to 512 input lines and output lines should be straightforward and result in aggregate throughputs of greater than 1 Terabits per second. Scaling to an ATM switch size of 1024×1024 is considered within the present technology, and the architecture of the present invention is believed to be extensible even further as the speed of commercially available components increases and as new, faster technologies are developed.

Referring now to FIG. 4, a specific embodiment of an ATM switch 10A is shown. In this specific embodiment ATM switch 10A has two hundred fifty six input interfaces 12<sub>0</sub>-12<sub>255</sub> which are connected to two hundred fifty-six ATM input lines In<sub>0</sub>-In<sub>255</sub>. The outputs of the input interfaces are connected to the input ports 17<sub>0</sub>-17<sub>N-1</sub> of the switch fabric 14A. The switch fabric 14A contains a total of sixty-four 16×16 crossbar switches 15<sub>0</sub>-15<sub>63</sub> which are partitioned into four pipes 18<sub>0</sub>-18<sub>3</sub>. The fanout F is equal to four which if the number of output ports=FN results in 1024 output ports 19<sub>0</sub>-19<sub>1023</sub>. The output ports 19<sub>0</sub>-19<sub>1023</sub> are respectively connected to the inputs of sixteen 64×16 output packet modules 16<sub>0</sub>-16<sub>15</sub>. The sixteen 64×16 output packet modules are connected to two hundred fifty six outputs Out<sub>0</sub>-Out<sub>255</sub>. Those skilled in the art will recognize that other combinations of components could have been used, for example thirty two 32×8 output modules could have been used instead of the 64×16 output modules shown in FIG. 4.

ATM switch 10A also has a controller 20 which has the tasks of hunting and finding an available pipe through the switch fabric 14A for each ATM packet. The controller 20 uses the fact that the switch fabric 14A is partitioned into four pipes to break the pipe hunting tasks into four parallel pipe hunting tasks that are each temporally shifted by an acceptable amount. Details of one embodiment of such a controller 20 are shown in FIG. 5.

For the 0.640 Terabits per second, N=256 embodiment mentioned previously and shown in FIGS. 4 and 5, the controller 20 may be contained on approximately eight printed circuit boards. Controller 20 would accept up to 256 sixteen-bit request vectors from up to 256 line input interfaces 12<sub>0</sub>-12<sub>255</sub> and perform path hunts on each of these request vectors within each 176 nanosecond. ATM cell interval to create the 1024 sixteen-bit connect vectors used to establish connections within the switch fabric 14A. This requires that controller 20 operate with a processor clock rate of at least 46 Mbps. This moderate clock rate permits the logic within the controller 20 to be implemented with off-the-shelf CMOS EPLD's or similar devices, thus making the cost of the controller 20 (in large quantities) very reasonable.

The movement of request vectors from the input interfaces 12<sub>0</sub>-12<sub>255</sub> to the controller 20 and the movement of connect vectors from the controller 20 to the crossbar switches 15<sub>0</sub>-15<sub>63</sub> of the switch fabric 14A is a challenging task, because large amounts of control information must be transported every 176 nano seconds ATM cell interval. For example, in an ATM switch containing 256 input interfaces, 256 16-bit request vectors must be transported to the controller 20 every 176 nano seconds, leading to an aggregate bandwidth of 23 Gigabits per second between the input interfaces sub-system and the controller 20 sub-system. In addition, 1024 16-bit connect vectors must be transported to the switch fabric 14A every 176 nano seconds to control the crossbars switches 15<sub>0</sub>-15<sub>63</sub>. This requires an aggregate bandwidth of 93 Gigabits per second between the controller 20 sub-system and the switch fabric 14A sub-system. This 93 Gigabits per second connect vector information can be compressed into 29 Gigabits per second (given that only one input can be routed to an output during each ATM cell interval) by standard compression techniques. However, since this control information should be delivered with high reliability, all of the control connections or control links between these sub-systems should be dually redundant (not shown in FIG. 4), so there is actually 46 Gigabits per second of data moving between the input interfaces cards and the controller 20 and 58 Gigabits per second of data moving between the controller 20 and the switch fabric 14A. Preferably, high-speed serial links 22 will be used to transmit this control information. For such a case, input interfaces 12<sub>0</sub>-12<sub>255</sub> would be grouped by fours such that only sixty-four serial links would be required to move request vectors from the input interfaces 12<sub>0</sub>-12<sub>255</sub> to the controller 20, and 128 serial links would be required to move the resulting connect vectors from the controller 20 to the pipes 18<sub>0</sub>-18<sub>3</sub> (assuming the aforementioned data compression techniques are applied to the connect vectors).

While the use of out-of-band control techniques does require the additional hardware cost of these high-speed serial control links 22, these links 22 cause very little increase the overall system hardware cost. Considering that the 256-input ATM switch 10A of FIGS. 4 and 5 already has 1024 high-speed serial links required to route ATM cells between the input interfaces 12<sub>0</sub>-12<sub>255</sub> and the switch fabric 14A (when the fanout of four is included) and 1024

more high-speed serial links are used to route ATM cells from the switch fabric outputs  $19_0-19_{1023}$  to the output packet modules  $16_0-16_{15}$ . Thus, the addition of the 192 serial links 22 for routing of the control information increases the total number of high-speed serial links within the system by merely nine percent.

Applying the calculations of Yeh et al. from the article "The Knockout Switch" the ATM cell loss probability of the ATM switch 10A shown in FIGS. 4 and 5 is  $4.34 \times 10^{-3}$ , assuming that the connections of the inputs is symmetrical and not independent as set forth in our co-pending application entitled "TERABIT PER SECOND DISTRIBUTION NETWORK". This cell loss probability falls short of the acceptable ATM cell loss probability of less than  $1 \times 10^{-12}$  mentioned previously.

To reduce the ATM cell loss probabilities, controller 20 applies a temporal spreading technique known as rolling, which provides many statistical advantages. Rolling involves and fulfills three fundamental goals that are aimed at providing more evenly distributed traffic loads. These goals are: (1) spatially distribute the traffic evenly across all pipes  $18_0-18_3$ , so that one pipe will only carry its proportional fraction of the traffic load, (2) spatially distribute the traffic evenly across all of the  $16 \times 16$  crossbar switches  $15_0-15_{63}$  within each pipe  $18_0-18_3$ , so that each of the crossbar switches is equally loaded, and (3) temporally distribute the traffic that arrives in a given ATM cell period across two ATM cell periods so that the traffic load can be effectively decreased in an occasional ATM cell period when an unusually high volume of traffic exists and is destined for a particular output packet module. This effective lowering of the traffic load is accomplished by delaying some of the ATM cells arriving during a congested ATM cell interval. The cells are delayed until the next consecutive ATM cell interval when the traffic load competing for the popular resources, i.e. connections to popular output packet modules, will most likely be lower, so the delayed cells should have a higher probability of being routed in the next ATM cell interval. Since the switch fabric 14A is memoryless, the ATM cells that must wait for the next ATM cell interval are stored in their respective input interfaces  $12_0-12_{255}$ .

In addition to satisfying these three fundamental goals of packet traffic control to distribute the load, rolling also satisfies two further very important ATM system goals. First, goal (4) is that the ATM switch 10A must guarantee that ATM cell ordering can be simply maintained when an ATM stream is re-constructed at an output packet module  $16_0-16_{15}$ , even if rolling causes some of the ATM cells within the stream to be delayed differently than others. Secondly, goal (5) is that rolling must also guarantee that the controller 20 will attempt to route every ATM cell through each of the four paths to its desired output packet module, but each of the successive path hunt attempts must occur in a more lightly-loaded  $16 \times 16$  crossbar switch so that the first attempt occurs in a  $16 \times 16$  crossbar switch with many previously-routed ATM cells (and very few available paths to output packet modules) while the fourth and final path hunt attempt occurs in an  $16 \times 16$  crossbar switch that is virtually empty (thereby providing many available paths to output packet modules). The rolling technique is similar to spatial path hunt techniques that pack as many calls as possible in one portion of a spatial network, which by forcing near 100% occupancy in parts of a system results in the remainder of the calls having a very high probability of being successfully routed through the remainder of the system if usage is below 100%. Thus, rolling in its fourth

and final path hunt attempt provides a very high probability of an ATM cell successfully being routed. Goal (5), by packing many ATM cells in one portion of the network, superficially seems to conflict with goal (1) that requires the traffic be spatially distributed across the network. However, as will be explained below, temporal spreading provided by the rolling technique permits the network to simultaneously satisfy both goals (1) and (5).

Assuming that each of the 256 input ports  $17_0-17_{N-1}$  of FIG. 4 has an ATM cell that needs to be routed through the distribution network, and assuming that the switch fabric 14A is composed of four pipes  $18_0-18_3$ , then the out-of-band controller 20 may be required to perform  $256 \times 4 = 1024$  unique path hunts for the ATM cells before the cells can be routed. To distribute the ATM cells evenly across all four pipes, the 256 ATM cells requesting connections, the rolling technique divides the requests into four groups of equal size. The first group will have path hunts performed for its ATM cells in pipe  $18_0$ , first, then in pipe  $18_1$ , then in pipe  $18_2$ , and finally in pipe  $18_3$ . The second group will have path hunts performed for its ATM cells in pipe  $18_1$ , first, then in pipe  $18_2$ , then in pipe  $18_3$ , and finally in pipe  $18_0$ . The third group will have path hunts performed for its ATM cells in pipe  $18_2$ , first, then in pipe  $18_3$ , then in pipe  $18_0$ , and finally in pipe  $18_1$ . The fourth group will have path hunts performed for its ATM cells in pipe  $18_3$ , first, then in pipe  $18_0$ , then in pipe  $18_1$ , and finally in pipe  $18_2$ . This ring-like ordering of the path hunts guarantees that the routed ATM cells are distributed evenly across all four pipes. In addition, if the ATM cells within each of the four equally sized groups are selected such that the ATM cells within a single group can be routed into exactly four of the 16 inputs on any  $16 \times 16$  crossbar switch, then the routed ATM cells will also be evenly distributed across all of the  $16 \times 16$  crossbar switches.

Referring now to FIGS. 5 and 6, a timing diagram for a rolling technique according to the present invention is described. To satisfy goals (1), (2), and (5) simultaneously, the out-of-band controller 20 uses the time delay/time distribution described in goal (3), and these ATM cell delays required by goal (3) must be provided during each ATM cell interval. In all cases, when a group of ATM cells is passed around the ring-like structure of controller 20 from pipe  $18_3$  to pipe  $18_0$ , the controller 20 re-assigns the cells to the next ATM cell interval (period) which requires that the ATM cells be delayed by one cell period. Because of this re-assignment and delay, each cell group encounters a very lightly-loaded set of  $16 \times 16$  crossbar switches for its fourth and final path hunt. An additional advantage of this rolling technique using re-assignment and delay of ATM cell intervals is that it also allows more than 64 simultaneously arriving ATM cells to be routed through the switch fabric 14A to any single output packet module  $16_0-16_{15}$  (even though there are only 64 connections or links from the switch fabric 14A to each output packet module  $16_0-16_{15}$ ). This is occurs with the rolling technique because all of the ATM cells do not need to be routed during the same ATM cell interval. Thus, the rolling technique when used in the out-of-band controller 20 results in extremely low cell loss probabilities both within the switch fabric 14A and the output modules  $16_0-16_{15}$ , even during a transient cell interval that has an extraordinarily high traffic load.

The one ATM cell period delays incurred by some of the ATM cells as they are routed through the switch fabric 14A would normally lead to the conclusion that there would be difficulties in satisfying goal (4) of maintaining proper cell ordering. However, the ring-like ordering of the path hunts within the out-of-band controller 20 guarantees that delayed

cells in a stream of ATM cells will always be routed through lower-numbered pipes than non-delayed cells (where pipe 18<sub>0</sub> is the lowest-numbered pipe and pipe 18<sub>3</sub> is the highest-numbered pipe). This information, coupled with the fact that ATM cells are delayed by at most one cell period, ensures that proper cell ordering will be maintained if the cells are extracted from the switch fabric 14A and loaded into first-in-first-out queues 174<sub>0</sub>-174<sub>15</sub> (shown in FIG. 7) of each output module of the output modules 16<sub>0</sub>-16<sub>15</sub> in the order of the lowest numbered pipe to the highest numbered pipe: pipe 18<sub>0</sub>, pipe 18<sub>1</sub>, pipe 18<sub>2</sub>, and pipe 18<sub>3</sub>.

Referring now to FIG. 7, the output module 16<sub>0</sub> (and the fifteen other output modules 16<sub>1</sub>-16<sub>15</sub>) may be a 64×16 embodiment of the concentrator described in U.S. Pat. No. 5,412,646, entitled "ASYNCHRONOUS TRANSFER MODE SWITCH ARCHITECTURE", filed May 13, 1994, by Cyr et al. and commonly assigned to the assignee of the present invention, which is hereby incorporated by reference. The output module 16<sub>0</sub> in FIG. 7 is a specific case of the generalized concentrator shown in FIG. 4 of the above-referenced patent application of Cyr et al. Since the output modules 16<sub>0</sub>-16<sub>15</sub> are well described in the above-referenced application, in the interest of brevity they will not be further described here.

To provide a better understanding of the equation of the rolling technique, a real-life analogy will be described with respect to FIG. 8, which is a plan view of an amusement park system 500. Consider the problem of transporting a large number of people from amusement park parking lots 511, 512, 513, or 514 to the amusement park 520 using trams to shuttle the people between the two points. Tram system 530 is composed of four tram shuttle trains each with a predetermined route, which is analogous to the four pipes of switch fabric 14A. Each tram shuttle train contains sixteen cars (representing the 16×16 crossbar switches within a particular pipe), and each shuttle car is equipped with sixteen seats (representing the output links emanating from a single 16×16 crossbar switch). In this analogy, each customer (representing an ATM cell) arrives in one of four parking lots 511, 512, 513, or 514 surrounding the amusement park 520. As a result, each customer is instantly placed in one of four groups, and since the parking lots 511-514 are the same size, each group contains an equal number of customers on the average. The customers in any single parking lot 511, 512, 513, or 514 must then divide up and stand in one of sixteen lines, where each line is associated with a respective car of the tram shuttle train. The amusement park 520 is sub-divided into sixteen different theme areas (The Past Land, The Future Land, etc.), and each of the sixteen seats of a particular tram car is labeled with the theme area to which that seat's occupant will be given admission. Before arriving in the parking lot, each customer must randomly chose one of the sixteen theme areas (representing the sixteen output packet modules 16<sub>0</sub>-16<sub>15</sub>) where he or she wishes to spend the day. Customers must then find an available seat associated with their desired theme area on one of the four trams that passes by the loading area 531, 532, 533, or 534 of their parking lot. If a customer has not found an available seat after four trams have passed by, then he or she is not permitted to enter the amusement park during that day (This harsh condition represents the loss of an ATM cell due to blocking in all four pipes of the distribution network, a small but finite possibility).

The first tram that stops at the loading area that the customer can try has already visited three other parking lot loading areas, so the customer's pre-specified seat may be

full. However, if the customer does find his or her seat to be vacant on that tram, then the tram will deliver him or her straight to the amusement park 520. If the customer fails to get on the first tram, he or she must wait and try the second tram which has already visited two other parking lot loading areas. If the customer is successful at finding his or her pre-specified seat on the second tram, that tram will deliver the customer to the amusement park 520 after one more parking lot stop. If the customer fails to get on the first tram and the second tram, then he or she must wait and try the third tram which has only visited one other parking lot loading area. If the customer is successful at finding his or her seat on the third tram, that tram will deliver him or her to the amusement park 520 after two additional parking lot stops. If the customer fails to get on any of the first three trams, then the customer must wait and try the fourth and final tram. Fortunately, this tram has not visited any parking lots yet, so the arriving tram is empty, and the customer's seat will be taken only if another customer in his/her parking lot line is also trying for the same seat. The system 530 satisfies goal (5), because each of the successively arriving trams is more lightly-loaded than the previous one. Thus, a controller 20 rolling ATM cells indeed can fulfill goals (1), (2), and (5).

The rolling technique if used by itself improves the ATM cell loss probability of ATM switch 10A from  $4.34 \times 10^{-3}$  to approximately  $10^{-11}$ . Using the analysis techniques of the article "A Growable Packet Switch Architecture" the cell loss probabilities for an ATM switch 10A that has independent connections to inputs of the switch fabric 14 according to Galois field theory and also has an out-of-band controller 20 that incorporates rolling techniques can be analytically modeled and calculated. Each of the 16×16 crossbar switches in pipe 18<sub>0</sub> receives an offered traffic load equal to  $R_a = R_L/4 + R_{res}$ , where  $R_{res}$  is defined to be the fraction of the 16 inputs to a 16×16 crossbar switch that are blocked in pipe 18<sub>3</sub>, and routed to pipe 18<sub>0</sub> for a re-attempt. For a first attempt at solving for the cell loss probability, let us assume that  $R_{res} = R_L/16$ . Thus, the cell loss probability of a single 16×16 crossbar switch in pipe 18<sub>0</sub> can be determined using the equation of Eng et al.

$$P(\text{cell loss}) =$$

$$[1 - m/(nR_L)] \left[ 1 - \sum_{k=0}^m \{ (nR_L)^k e^{-(nR_L)} \} / k! \right] + (nR_L)^m e^{-(nR_L)}/m!$$

where  $m=1$ ,  $n=1$ , and the switch loading is given by  $R_a = R_L/4 + R_L/16$ . Using these assignments, the resulting cell loss probability for a fully-loaded ( $R_L = 1.0$ ) pipe 18<sub>0</sub> 16×16 crossbar switch can be calculated to be:

$$P(\text{cell loss in pipe } 18_0) = 1.3 \times 10^{-4}$$

Thus, the fraction of the 16 inputs to a 16×16 crossbar that are passed to the second pipe after the first attempt is given by:

$$f_1 = R_a \times P(\text{cell loss in pipe } 18_0) = (3.13 \times 10^{-4}) \times (1.3 \times 10^{-4}) = 4.06 \times 10^{-7}$$

By symmetry, this should have also been the same as the fraction of inputs that are passed from pipe 18<sub>3</sub> to pipe 18<sub>0</sub>, so the residue assumption of  $R_L/16 = 0.062$  above was incorrect. By refining this assumption and performing a second attempt, and now assuming that  $R_{res} = R_L/32$ . Thus, the cell

loss probability of a single  $16 \times 16$  crossbar switch in pipe 18<sub>0</sub> can be determined again using the equation of Eng et al., where  $m=1$ ,  $n=1$ , and the switch loading is given by  $R_a=R_L/4+R_L/32$ . Using these assignments, the resulting cell loss probability for a fully-loaded ( $R_L=1.0$ ) pipe 18<sub>0</sub>  $16 \times 16$  crossbar switch is calculated to be:

$$P(\text{cell loss in pipe } 18_0) = 1.2 \times 10^{-3}$$

Thus, the fraction of the 16 inputs to a  $16 \times 16$  crossbar that are passed to the second pipe after the first attempt is given by:

$$f_{1-2} = R_a \times P(\text{cell loss in pipe } 18_0) = (2.81 \times 10^{-4}) \times (1.2 \times 10^{-3}) = 3.37 \times 10^{-7}$$

This calculation result is very close to the assumed value of  $R_{\text{res}}=R_L/32=3.13 \times 10^{-2}$ , so the assumption is considered to be satisfactory. The blocked cells are sent to pipe 18<sub>1</sub> for subsequent path hunting, and they encounter a negligible number of ATM cells from previous attempts. Thus, the  $16 \times 16$  crossbar switch in pipe 18<sub>1</sub> can be modeled for analysis as a growable packet switch, with  $m=1$ ,  $n=1$ , and  $R_a=f_{1-2}$ , and the resulting cell loss probability of this model is  $1.4 \times 10^{-21}$ . The fraction of the 16 inputs to the  $16 \times 16$  crossbar in pipe 18<sub>1</sub> that are passed to the pipe 18<sub>2</sub> is  $4.2 \times 10^{-4}$ . Similar arguments can be used to show that the resulting cell loss probability for cells entering pipe 18<sub>2</sub> is  $1.9 \times 10^{-4}$ , and the resulting fraction of the 16 inputs to a  $16 \times 16$  crossbar passed to pipe 18<sub>3</sub> is  $7.9 \times 10^{-8}$ . The resulting ATM cell loss probability in pipe 18<sub>3</sub> is  $3.7 \times 10^{-6}$ , and the fraction of the 16 inputs to a  $16 \times 16$  crossbar not routed in pipe 18<sub>3</sub> (and therefore not routed in all four pipe attempts) is  $2.9 \times 10^{-15}$ . Thus, through the use of the rolling techniques within the out-of-band controller 20, the ATM cell loss probability of an ATM switch 10A with independent connections at the inputs of its switch fabric 14A can be decreased from an unacceptable value of  $1.47 \times 10^{-6}$  to an acceptable value of  $2.9 \times 10^{-15}$ .

A preference technique may be used in conjunction with the rolling technique described above to decrease the cell loss probability of an ATM switch 10A even further. Referring back to FIG. 8 and the amusement park analogy, some form of arbitration was required at the tram loading areas to determine which of the customers in the line will be given a particular seat on the tram when more than one customer is requesting the same seat. Similarly, the out-of-band controller 20 must provide an arbitration scheme for selecting which of the arriving ATM cells will be assigned a particular link whenever two or more cells request access to the same link. The arbitration scheme used can have an advantageous effect on the ATM cell loss probabilities.

One possible arbitration scheme is a random scheme to determine which of the ATM cells is assigned the link. The random selection scheme is the scheme assumed for the analysis of the rolling technique presented above. However, other arbitration schemes are possible, and one particular arbitration scheme that has advantageous results is called the preference scheme. The preference arbitration scheme assigns a preference weight to each of the ATM cells in a particular grouping. ATM cells with higher preference weights are given precedence over ATM cells with lower preference weights whenever two or more cells request access to the same link. As a result, an effective hierarchy is created within the groupings of ATM cells.

The creation of a hierarchy may superficially seem to produce undesirable characteristics within the switch fabric

14A, because customers with high preference weights will be offered better service than customers with low preference weights. In fact, the one customer with the highest preference weight within each group can never have his or her ATM cell blocked by another customer's ATM cell. Although this may seem unfair, a detailed analysis of the effects of imposing this hierarchy indicates that it actually leads to improved performance, i.e. lower cell loss probabilities, for all customers—even for the customer at the bottom of the hierarchy with the very lowest preference weight.

The results of this analysis are summarized in FIG. 9, where the probability of loss of an ATM cell; i.e., the probability of a cell not being assigned to an available path, is shown as a function of the number of path hunts that were attempted in different pipes by the out-of-band controller 20. In this analysis, it was assumed that the group sizes were four—i.e., up to four ATM cells could simultaneously compete for access to the same link. As a result, four different preference weights were assigned to create a hierarchy for the four input ports associated with each group. The preference weight associated with a particular input port is assumed to be a fixed constant that does not vary with time. The resulting plots 901, 902, 903 and 904 in FIG. 9 indicate that the cell loss probability decreases as more path hunts in more pipes are performed, but it also shows that the inputs with the lower preference weights 903, 904 have higher cell loss probabilities than the inputs with higher preference weights 901, 902, as might be expected. Super-imposed on these plots is a similar plot 910 which indicates the probability of not being served when a random selection arbitration scheme is used instead of the hierarchy arbitration scheme. The surprising and unexpected results are that after path hunt attempts in four different pipes, the random selection arbitration scheme produces cell loss probabilities which are higher than the average of the cell loss probabilities for the hierarchy arbitration scheme. In fact, the plot 910 of the random selection arbitration scheme shows an average cell loss probabilities for all of the input ports which are notably higher than the plots 903 and 904 which are the average cell loss probabilities for even the input ports with the lowest preference weights within the hierarchy arbitration scheme. This phenomenon can be explained by the fact that after three sets of path hunts in three different pipes, the distribution of ATM cell requests entering the fourth pipe is very different depending on whether the random or preferences arbitration scheme is used. In the random selection arbitration scheme, there is a small but equal probability that all of the ATM cells are requesting a path. However, in the hierarchy arbitration scheme, most of the ATM cells with higher preference weights will be requesting a path with a probability of practically zero, while the ATM cell with the lowest preference weight will be requesting a path with a sizable probability, because that particular ATM cell may have been denied access to links in all three of its previous path hunt attempts. However, a single request arriving with a high probability at the fourth and last path hunter in the controller will lead to more routed ATM cells than many requests arriving with low probability, because the single request can always be satisfied since contention for an output link will never occur.

As a result, it seems apparent from the plots in FIG. 9 that by assigning preference weights to the input ports and by using a hierarchy arbitration method to resolve link contention and route paths in the out-of-band controller, the worst-case cell loss probability of the switch fabric 14A can be decreased from  $2.9 \times 10^{-15}$  that was achieved by the intro-

duction of the rolling technique to an even lower value of  $2.4 \times 10^{-16}$ . It is worth noting that input ports that are assigned higher preference weights will encounter even lower cell loss probabilities as indicated in FIG. 9.

Referring back to FIG. 5, in order to provide a physical embodiment of the rolling and preference methods, the ATM switch 10A is segmented into four basic sub-systems. These four sub-groups consist of the input interfaces  $12_0-12_{255}$ , the output modules  $16_0-16_{15}$ , the switch fabric 14A, and the out-of-band controller 20.

The input interfaces  $12_0-12_{255}$  within the network provide the necessary interfaces between the incoming transmission links and the links connected to the switch fabric 14A and the out-of-band controller 20. As a result, the input interfaces  $12_0-12_{255}$  must provide a termination for the input transmission line. For example, if the input transmission line is a SONET link, then the input interface must provide for clock recovery, link error detection, SONET pointer processing and frame delineation, ATM cell extraction, and an elastic storage function to synchronize the arriving ATM cells to the system clock within the distribution network. The extracted ATM cells are then loaded into a FIFO buffer of the input interface. The input interface must also read ATM cells from the FIFO buffer and extract the ATM header from the cell. The VPI/VCI field of each ATM header is then used as an address into a translation table located on the input interface. The output of the translation table provides a new VPI/VCI field and the address of the output packet module to which the ATM cell is to be routed. The new VPI/VCI field is written into the ATM cell as a replacement for the old VPI/VCI field, while the output module address is routed as a request vector to the out-of-band controller 20 for the controller fabric 14A. Since the amount of processing time required by the out-of-band controller 20 is a fixed value, the input interface simply holds the ATM cell in a buffer until the out-of-band controller 20 has completed its path hunt and has relayed the results into the switch fabric 14A. Once the switch fabric 14A is loaded with the new switch settings to appropriately route the ATM cell, the input interface can inject the ATM cell into the switch fabric 14A and it will be automatically routed through the switch fabric 14A to its desired output module  $16_0-16_{15}$ . It should be noted that each input interface  $12_0-12_{255}$  actually is provided with one link to each of the four pipes  $18_0-18_3$  of the switch fabric 14A. In addition, the use of rolling (i.e. temporal spreading) within the switch fabric 14A may require a copy of the ATM cell to be injected into each of the four links during any one of two consecutive ATM cell intervals. As a result, the timing within the input interfaces  $12_0-12_{255}$  must be tightly coupled and synchronized to the timing of the rest of the sub-systems within the ATM switch 10A.

Each of the two hundred fifty six input interfaces  $12_0-12_{255}$  in FIG. 5 are numbered with an address ranging from 0 to 255, but each input interface is also assigned an alias address given by a letter between A and P. These alias addresses are used to identify which input port the input interfaces will connect to within the switch fabric 14A. The actual set of four crossbar switches to which a particular input interface is connected is determined by the Galois field techniques that were described previously. These techniques guarantee independence between all of the inputs on any  $16 \times 16$  crossbar switch of any pipe.

Each of the sixteen output modules  $16_0-16_{15}$  in FIG. 5 is labeled with addresses ranging from AA to PP, and each output module performs an important function within the ATM switch 10A. Each of the output modules  $16_0-16_{15}$

within FIG. 5 provides terminations for a respective set of sixty-four links emanating from the switch fabric 14A. Each output module  $16_0-16_{15}$  also provides two basic functions: it provides a small degree of space switching to route each ATM cell arriving on one of the sixty-four inputs to the desired one of the sixteen output ports, and it provides buffering of ATM cells to handle the problems associated with multiple packets that are simultaneously destined for the same output  $Out_0-Out_{255}$ .

There are many ways for these two functions to be implemented. The most straight-forward approach would probably construct a shared memory switch that could perform sixty-four memory writes and sixteen memory reads within an ATM cell interval (176 nano seconds). The memory could then be treated as sixteen disjoint linked lists (one for each output  $Out_0-Out_{255}$ ) along with a seventeenth linked list containing idle memory locations. Although simple, this approach requires eighty memory accesses every 176 nano seconds, so it would demand memories with 2.2 nano seconds access times. An alternate approach would split each  $64 \times 16$  output module  $16_0-16_{15}$  into a  $64 \times 16$  concentrator and a  $16 \times 16$  shared memory switch. The concentrator would be a memory system that provides for sixty-four writes and sixteen reads every ATM cell interval, but the memory size could be small (and memory speeds could be fast) since the buffering required for output contention problems is not provided in this memory. In addition, the  $64 \times 16$  concentrator could be implemented as a single linked list spread out across sixty-four distinct memory chips. As a result, each memory chip would require only one write and up to sixteen reads for every ATM cell interval. The  $16 \times 16$  shared memory switch only performs thirty-two memory accesses every ATM cell interval, so slower (and larger) memories could be used, and the buffering for output contention problems could be provided in this shared memory portion of the output module. Thus, this latter arrangement is the more practical alternative for an output module.

The switch fabric 14A is essentially a group of small circuit switches that provide the required connectivity between the input interfaces and the output modules in response to the control signals generated by the out-of-band controller 20. In the embodiment of the ATM switch 10A shown in FIG. 5, the switch fabric 14A is composed of sixty-four  $16 \times 16$  crossbar switches, where disjoint groups of sixteen switches comprise a pipe. The four pipes are labeled pipe  $18_0$ , pipe  $18_1$ , pipe  $18_2$ , and pipe  $18_3$ , and the sixteen  $16 \times 16$  crossbar switches within a given pipe are labeled switch 0-15. The crossbar switches must be capable of receiving the control signals generated by the out-of-band controller 20 and must re-configure all of the switch settings during a guard-band interval between consecutive ATM cells. Each  $16 \times 16$  crossbar switch supports sixteen inputs labeled input A through input P, and each  $16 \times 16$  crossbar switch also supports sixteen outputs labeled output AA to output PP. It was noted above that each input interface connects to a different  $16 \times 16$  crossbar in each of the four pipes  $18_0-18_3$ , but it should now be noted that an input interface that connects to input X in pipe  $18_0$  is required to be connected to input X in the other three pipes  $18_1-18_3$ , as well, where X is an element of the set {A,B, . . . ,P}. The actual connections between the input interfaces  $12_0-12_{255}$  and the crossbar switches within the switch fabric 14A are determined using Galois field theory techniques that were referenced above. These techniques guarantee independence between input ports for routing within switches in each pipe of the switch fabric 14A. FIG. 5 also illustrates that output

YY from each of the sixty-four crossbar switches is routed to one of the sixty-four inputs on the  $64 \times 16$  output module labeled YY, where YY is an element of the set {AA, BB, .., PP}.

The basic function of the out-of-band controller 20 for the switch fabric 14A is to determine through which of the four pipes 18<sub>0</sub>-18<sub>3</sub> a particular ATM cell may be routed. Once the out-of-band controller 20 has successfully determined a pipe through which the ATM cell is to be routed without being blocked, the task of setting up the path through the pipe is simple, because by the definition of a pipe, there will exist only one path within the pipe between the input port of the arriving ATM cell and the desired output module. As a result, the fundamental path hunting task of a switching network is essentially reduced to the simpler task of pipe hunting in the ATM switch 10A.

The out-of-band controller 20 still requires a large busy-idle table to identify the status of each of the intermediate (FN) links between the  $16 \times 16$  crossbar switches of the switch fabric 14A and the output modules 16<sub>0</sub>-16<sub>15</sub> as busy and unavailable or idle and available. However, this large busy-idle table may be sub-divided into many small busy-idle tables that the controller 20 can access in parallel, and thereby perform many pipe hunting operations in parallel. There are many ways to implement the controller 20 for a large switch having the general growable packet switch architecture. In the extreme case, four levels of parallelism may be applied to the architecture of the controller 20 to perform pipe hunting. One embodiment that uses three levels of parallelism will be described in detail, first and then a fourth level of parallelism for the controller 20 will be discussed.

The first level of parallelism is obtained by providing each of the four pipes 18<sub>0</sub>-18<sub>3</sub> with a respective pipe hunt controller 24<sub>0</sub>-24<sub>3</sub>. This level of parallelism allows pipe hunting to be carried out in all four pipe hunt controllers 24<sub>0</sub>-24<sub>3</sub> simultaneously. The second level of parallelism is obtained by providing switch controllers 26<sub>0</sub>-26<sub>63</sub>, with sixteen switch controllers within each pipe hunt controller 24<sub>0</sub>-24<sub>3</sub>. A unique switch controller 26<sub>0</sub>-26<sub>63</sub> is respectively associated with each of the  $16 \times 16$  switches within each pipe of the switch fabric 14A. As a result, pipe hunting operations can be carried out in parallel within all sixteen of the switch controllers of each pipe hunt controller 24<sub>0</sub>-24<sub>3</sub>. The third level of parallelism is obtained by permitting each of the switch controllers 26<sub>0</sub>-26<sub>63</sub> to perform parallel processing over all sixteen of the output links attached to its respective  $16 \times 16$  crossbar switch. Effectively, each of the switch controllers 26<sub>0</sub>-26<sub>63</sub> reads sixteen busy-idle bits from its busy-idle memory in parallel, performs parallel pipe hunting operations based on those sixteen bits, and then writes the sixteen resulting busy-idle bits into its respective busy-idle memory in parallel with the other busy-idle memories. A representative switch controller 26<sub>0</sub> of the sixty four switch controllers 26<sub>0</sub>-26<sub>63</sub> is shown in FIG. 10. The concurrent processing of sixteen busy-idle bits is accomplished by providing switch controller 26<sub>0</sub> sixteen unique link controllers AA-PP, each of the link controllers AA-PP is assigned the task of processing busy-idle bits for one intermediate link between its portion of the switch fabric 14A and its respective output modules. In the embodiment shown in FIG. 10, the large busy-idle memory required to control switch 10A has been divided into many single bit memories, busy-idle flip-flops, with each single bit, busy-idle memory being logically and physically associated with its respective link controller AA-PP.

The general data flow for request vectors generated by the input interfaces 12<sub>0</sub>-12<sub>255</sub> is shown in FIG. 5. For example,

input interface 12<sub>0</sub> in FIG. 5 routes its request vector to pipe hunting controller 24<sub>0</sub> where it is poked into the pipe hunting ring (i.e. controller 20), and the rolling scheme requires the request vector to be looped through pipe hunt controller 24<sub>1</sub>, pipe hunt controller 24<sub>2</sub>, and pipe hunt controller 24<sub>3</sub>, as it circulates around the ring. In general, each of the input interfaces 12<sub>0</sub>-12<sub>255</sub> produces one request vector, and each request vector will contain a number of bits equal to the number of output modules within the system. The request vector from a single input interface in FIG. 5 is thus a sixteen-bit data word, where each bit of the request vector points to one of the sixteen output modules. If an ATM cell within a input interface is requesting a connection to an output port on the i-th output module, then bit i within the request vector will be set to a logic "1" and all other bits within the request vector will be set to a logic "0". When the controller 20 receives this particular request vector from the input interface, it can then identify that a path is required between the source input interface and the i-th output module.

The entire sixteen-bit request vector from a input interface is routed via a respective control connection 21<sub>0</sub>-21<sub>255</sub> to one of the four pipe hunt controllers 24<sub>0</sub>-24<sub>3</sub>, and the controller 20 pokes the vector into one of the sixteen switch controllers associated with that particular pipe hunt controller. As shown in FIG. 10, the sixteen bits of the request vector are injected into a switch controller and are distributed across all sixteen of the link controllers within that particular switch controller. Each link controller is associated with a single link between the crossbar switches and the output modules, and it essentially processes one bit of the sixteen-bit request vector. This finite state machine circuitry that is associated with a single link controller consists of one flip-flop (the single-bit memory required to store the busy-idle bit associated with this link controller's link) and four logic gates (shown in FIG. 11). A state table description of the link controller operation is given in FIG. 12, where the state variable is defined by the busy-idle bit. The link controller hardware provides for one request vector input bit, designated request-in; one request vector output bit, designated request-out; and one connection vector output bit, designated connect. The request vector input bit is a logic "1" if the input desires a connection through the link associated with this link controller—otherwise, it is a logic "0". The request vector output bit is a logic "1" if the logic "1" input request vector bit was not satisfied by this particular link controller—otherwise, it is a logic "0". The connect vector output bit is a logic "1" if the logic "1" input request vector bit was satisfied by this particular link controller indicating the ATM cell will be routed to its desired output module through the link associated with this link controller—otherwise, it is a logic "0". The busy-idle flip-flop in FIG. 10 is reset to the logic "0" (idle) state at the beginning of each ATM cell slot, so the first request vector bit that enters the link controller with a logic "1" request is assigned the link (creating a logic "1" connect vector bit and a logic "0" output request vector bit) and sets the busy-idle flip-flop to the logic "1" (busy) state. Any subsequent request vector bits that enter the link controller during this particular ATM cell slot will be denied a connection through this link (forcing a logic "0" output on the connect vector bit and creating an output request vector bit that is identical to the input request vector bit). A time-lapsed view of several consecutive sixteen-bit request vectors passing through a single switch controller is shown in FIG. 12, along with the resulting states of the busy-idle bits stored within the switch controller. The resulting output request vectors and output

connect vectors illustrate the general operation of each of the pipe hunt controllers 24<sub>0</sub>-24<sub>3</sub>.

The use of rolling within the controller 20 requires a very precise temporal ordering of two fundamental events: poking and busy-idle flip-flop clearing. The timing diagram of FIG. 13 illustrates the synchronization and data flow that might be used for the logic within the controller 20. As indicated by the timing diagram, the flow of data around the ring of controller 20 is from pipe controller 24<sub>0</sub> to pipe controller 24<sub>1</sub>, to pipe controller 24<sub>2</sub>, to pipe controller 24<sub>3</sub>, and back to pipe controller 24<sub>0</sub>. Request vectors generated by input interfaces with alias addresses A, B, C, and D are poked into pipe controller 24<sub>0</sub>. Request vectors generated by input interfaces with alias addresses E, F, G, and H are poked into pipe controller 24<sub>1</sub>. Request vectors generated by input interfaces with alias addresses I, J, K, and L are poked into pipe controller 24<sub>2</sub>. Request vectors generated by input interfaces with alias addresses M, N, O, and P are poked into pipe controller 24<sub>3</sub>. The poking times and busy-idle bit clearing times take place at different moments within each of the pipe hunt controllers 24<sub>0</sub>-24<sub>3</sub>. From the point of view of any pipe controller, the request vector bits flow through the pipe controller in alphabetical order (A to P) if one ignores the busy-idle bit clearing times. This ordering guarantees that the aforementioned advantages of preferences will be realized within the controller 20, because the request vector generated from a input interface with alias address A will always be given precedence over the request vectors generated from input interfaces with alias addresses B, C, and D, etc.

The benefits derived from forced independence between the inputs on a particular 16×16 crossbar switch produce a slight increase in the complexity of the pipe hunter circuitry. Because of the independent connections between the input interfaces and the switch fabric 14A, which independence is assured by the use of Galois field theory, a request vector from a single input interface must be appropriately routed to several different switch controllers in each of the stages in the pipe hunting ring. The mixing nature of the Galois field theory generated connections requires each input interface 12<sub>0</sub>-12<sub>33</sub> to be connected to a different set of 16×16 crossbar switches within the switch fabric 14A, and as a consequence, it also requires request vectors generated on different input interfaces to be routed through entirely different sets of switch controllers within the controller 20. Since request vectors are time-multiplexed on links within the controller 20, all of the request vectors (within a particular ATM cell slot) that are expelled from a particular switch controller in one pipe hunter stage must (by definition) be routed to different switch controllers in the next pipe hunter stage. To provide this dynamic routing of the request vectors, each pipe hunt controller 24<sub>0</sub>, 24<sub>1</sub>, 24<sub>2</sub>, and 24<sub>3</sub> is connected to a respective small switching network 30<sub>0</sub>, 30<sub>1</sub>, 30<sub>2</sub>, and 30<sub>3</sub>, shown in FIG. 5. Alternatively, simple multiplexers may be used instead of switching networks 30<sub>0</sub>, 30<sub>1</sub>, 30<sub>2</sub>, and 30<sub>3</sub>, thereby greatly decreasing costs for the controller 20. Fortunately, the required configurations of these small switching networks 30<sub>0</sub>, 30<sub>1</sub>, 30<sub>2</sub>, and 30<sub>3</sub> (or multiplexers) are cyclic with a period equal to the ATM cell period, and the required configurations can be determined *a priori* and can therefore be "hard-coded" into the small switching networks (multiplexers) during the design of the circuitry of the controller 20.

As mentioned previously, ATM switch 10A shown in FIG. 5 might be scaled such that the number of input lines were 512, 1024 or even higher. For those size switches, assuming that the input lines are carrying 2.5 Gigabits per second data

rates, the aggregate throughput would be over 1.0 Terabits per second. For switches of that size, a fourth level of parallelism may be needed to provide sufficient processing power for the controller 20 to hunt for all the paths through all the pipes in real time. For ATM switches with 512 and 1024 input lines, the data rates on connections within their respective controllers are 204 Mbps and 386 Mbps, which is considerably higher than the 113 Mbps rate of the 256 input line version of ATM switch 10A.

The basic idea behind the fourth level of parallelism is a modification of the previously described controller 20 design which requires that request vectors be routed through the pipe hunter stages in parallel. In particular, all of the request vectors that are poked into a particular pipe are routed through the pipe hunter stages together, and these request vectors are said to comprise a poke group. In the embodiment shown in FIG. 5, this approach to the design of controller 20 creates four poke groups of sixteen-bit request vectors, so each poke group contains sixty-four bits. The four poke groups can be labeled with a concatenation of the four alias labels on the request vectors. As a result, the four poke groups for the re-designed pipe hunter of FIG. 5 are called ABCD, EFGH, IJKL, and MNOP. It is important to note that whenever a single sixty-four bit ABCD poke group is being routed through one of the switch controllers in pipe controller 24<sub>0</sub> of FIG. 5, there is also a sixty-four bit ABCD poke group being routed through each of the other fifteen switch controllers in pipe controller 24<sub>0</sub>. As a result, there are a total of 1024 request vector bits associated with sixteen ABCD poke groups that are being routed through pipe 18<sub>0</sub> at a single instant of time. The modified controller 20 processes the request vectors for all N input ports (by passing them through all four pipe hunt controllers 24<sub>0</sub>-24<sub>3</sub>) every eight clock cycles, and since this task must be completed within a single 176 nano seconds ATM cell interval, the required clock rate within the controller 20 is 46 Mbps regardless of the size (aggregate throughput) of the NxN ATM switch. As a result, since the controller 20 must perform eight processing steps (regardless of the network size), the process is said to be an O(1) path hunt algorithm. During the execution of this O(1) path hunt algorithm for the N=256 input ATM switch 10A of FIG. 5, the equivalent of 16,384 link controller path hunts and 16,384 link controller path hunt checks are performed every 176 nano seconds, so if each path hunt is considered to be an instruction execution and each path hunt check is considered to be an instruction execution, then the controller 20 can be viewed as a parallel processor capable of sustaining a 186 gigabits per second processing rate. The trade-off for maintaining a reasonable data rate in the controller 20 (regardless of size) is an increase in link controller logic complexity and an increase in signal connections passing between successive stages of the controller as the size is increases. ATM switch designs with aggregate throughputs in excess of 1 Terabits per second will require between 4096 and 32,768 signals (at 46 Mbps) to be routed between successive pipe controller stages.

In addition to increasing the number of signals between pipe controller stages, the use of parallelism within the controller 20 also requires a slight increase in the hardware requirements for each link controller, because each link controller must now support a parallel path hunt on four bits within the poke group. Even though a few additions are required to the hardware of the controller 20, the lower processing rate resulting from the next level of parallelism will typically justify minor increases in hardware.

In a large ATM system, a large percentage of the overall system cost will undoubtedly be concentrated in the output

packet modules, because the functionality within the output packet modules is relatively complex. In particular, each output packet module must provide a concentrator function that requires many FIFO-like memories, and the required access speed of these memories is directly proportional to the number of inputs and outputs of the concentrator. Each packet module must also provide a large shared buffer memory in order to accommodate packets that are temporarily blocked from their desired output port, and the required access speed of this memory is proportional to the number of outputs from the output packet module. The output packet module further provides a means of guiding packets to their desired output ports. Cell loss priorities and cell delay priorities must be managed within each output packet module as well as within the line cards and distribution network. It should be evident that larger output packet modules will make the task of providing all of these functions even more expensive and complex, so tradeoff techniques that help decrease the required size of the output packet module are beneficial.

In the system shown in FIG. 5, each of the output packet modules was designed to support sixty-four input links and sixteen output links. This size is directly related to the number of crossbar switches used in the distribution network, because each output packet module must have exactly one link routed to it from each of the crossbar switches in the distribution network. The number of outputs emanating from the output packet module is always  $1/F$  times the number of inputs to the output packet module, where  $F$  is the fanout provided by the distribution network. The fanout  $F$  is four in the network shown in FIG. 5.

As a result of these relationships, it should be evident that decreasing the number of crossbar switches within the distribution network will result in smaller output packet modules. However, to maintain low cell loss probabilities within the network, larger crossbar switches will be required so that there will still be a sufficient number of alternate paths through the distribution network. As an example, the sixty-four  $16 \times 16$  crossbar switches in FIG. 5 could be replaced by thirty-two  $32 \times 32$  crossbar switches, and the sixteen  $64 \times 16$  output packet modules of FIG. 5 could be replaced by thirty-two  $32 \times 8$  output packet modules. Another example of this approach occurs if the hardware of FIG. 5 is replaced by sixteen  $64 \times 64$  crossbar switches in the distribution network and sixty-four  $16 \times 4$  output packet modules. These examples indicate the tradeoffs between hardware in the distribution network and hardware in the output packet modules that can be made within the design of the ATM switch architecture of the present invention. Unfortunately, these trade-offs do not come without cost, because the designs that have larger crossbar switches in the distribution network (and smaller output packet modules) will typically produce lower overall system level costs, but they also produce degraded levels of system performance in cell loss probabilities. This degradation in the cell loss probability results from a decrease in the level of independence between inputs on the larger crossbar switches. Nevertheless, many designs still produce acceptable levels of cell loss without requiring the addition of a fifth active pipe in the distribution network. Simulations of the example having thirty-two  $32 \times 32$  crossbar switches and thirty-two  $32 \times 8$  output packet modules indicate a cell loss probability of  $1 \times 10^{-12}$ , which is only slightly higher than the cell loss probability for the network in FIG. 5.

The ATM switch architecture of the present invention which is based on out-of-band control techniques permits the use of global information with regard to the status of the

switching fabric 14A in many different ways. Likewise, the manner in which paths for arriving ATM cells are rapidly routed within the sections of the pipe hunter also provide opportunities for several features that would be very difficult in an architecture based on in-band/self-routing control techniques, as will be described below.

The ATM switch shown in FIG. 5 may also have multiple levels of cell loss priorities in order to provide different class of service levels to different customers. By defining a specific bit within the ATM cell header as the cell loss priority bit, two different cell loss priority service classes can be specified. Similarly, VPI/VCI field values or bits within the user payload might be used to identify even more class-of-service priority levels. Using either of these techniques, the ATM switch can be operated such that a higher priority cell always has a higher probability of being successfully routed through the network than a lower priority cell. Thus, if ATM cells are blocked and dropped, then the low priority cells should be the cells that are dropped. These cell loss priorities are completely different from the preference weights described previously with respect to switching network 14A. A high cell loss priority takes precedence over a high preference weight (described previously), but two cells with the same priority level should follow the arbitration rules defined by the ordering associated with the precedence weights.

The pipe hunter architecture has to be modified slightly from its operation shown in FIG. 14, in order to provide multiple cell loss priorities. This is easily accomplished because of the natural temporal ordering impressed on the request vectors as they pass through the pipe hunt controllers. This ordering puts an inherent level of priority to each of the request vectors. Thus, the non-parallel approach, shown in the timing diagram of FIG. 14, is modified to permit two different priority level groups of request vectors  $R'$  and  $R''$  to be passed through the pipe hunter in order to implement multiple priority levels. The request vectors associated with the high-priority ATM cells each is identified by a  $R'$ , and the request vectors associated with the low-priority ATM cells each is identified by a  $R''$ . The timing of the higher priority group of request vectors and the lower priority group of request vectors is illustrated in FIG. 15.

In FIG. 15, the request vectors have subscripts ABCD, y; EFGH, y; IJKL, y; and MNOP, y. The ABCD, EFGH, IJKL, and MNOP subscripts represent request vectors arriving in line cards with aliases ABCD, EFGH, IJKL, and MNOP respectively. The subscript y represents the cell time slot during which the ATM cells desire routing. In FIG. 15, these time slots are shown relative to an index time slot i. Along the left edge, each time slot is divided into twelve periods 0-11. The four columns to the right of the time column represent pipe controller 240, pipe controller 241, pipe controller 242 and pipe controller 243, respectively. The busy-idle bit vector for pipe p is for an ATM cell in time slot y is represented by  $B_{p,y}$ . Thus, for time slot i and pipe 0, i.e., pipe 240, the busy-idle bit vector is represented by  $B_{0,i}$ . The shading indicates the specific time periods when request vectors are poked into the pipe controllers. The lack of shading indicates that those request vectors arrived at that pipe controller via circulation around the circularly connected pipe controller described previously. For example, at time i, 6 and i, 7 request vectors  $R'_{ABCD,i}$  and  $R''_{ABCD,i}$  are respectively poked into pipe controller 0 and logically ANDED with  $B_{0,i}$  and  $B_{0,i+1}$  sequentially. If the busy-idle bits of vector  $B_{0,i}$  indicate that a requested path is available in pipe 0, then the path is assigned and the busy idle bit representing that path and the request vector requesting that

path are changed to reflect that the path is committed to fulfill a request in the present time period. At time i, 7 and i, 8 these request vectors (changed to reflect any fulfilled requests) are routed to pipe controller 1. At time i, 8 and i, 9 these request vectors (again changed to reflect any fulfilled requests) are routed to pipe controller 2. At time i, 9 and i, 10 these request vectors (again changed to reflect any fulfilled requests) are routed to pipe controller 3. At this time in pipe controller 3, request vectors  $R'_{ABCD,i}$  and  $R''_{ABCD,i}$  have the first two chances at selecting paths from a newly cleared pipe controller, the clearing of this pipe is represented by \*3. Thus, vector  $R'_{ABCD,i}$  should find paths for all its unfulfilled requests unless they self-conflict. These request vectors would continue to circulate, but at time i, 11 as these request vectors would return to pipe controller 0 the pipe controller is cleared (represented by \*0) and the request vectors  $R'_{ABCD,i}$  and  $R''_{ABCD,i}$  are discarded. As shown in FIG. 15, the other portions of the controller work in the same manner.

It should be noted that the inclusion of cell loss priorities techniques does not (and should not) detract from the desirable attributes of routing. In particular, each of the successive path hunt attempts for a particular ATM cell still occurs in a more lightly-loaded crossbar switch.

If a non-parallel controller approach (as shown in FIG. 5) to process two priority level request vectors, a data rate speed-up to permit all of the request vectors to be routed through the controller within the 176 nsec ATM cell interval. If this is deemed undesirable, then a parallel approach that routes the high-priority request vectors and low-priority request vectors through the controller simultaneously could also be used, but more signals are required between consecutive pipe controller stages and more complicated processing logic is required within the link controllers. Nevertheless, the required ordering needed to ensure that high-priority cells are routed before low-priority cells can still be maintained if the link controller hardware is designed with a hard-wired priority ordering.

Analytical modeling has shown that the distribution network 14A works well with multiple cell loss priority levels. For example, if the fully-loaded network of FIG. 5 had four priority levels (highest priority=level 0 and lowest priority=level 3) and has input traffic that is equally distributed across all four priority levels, then the distribution network 14A will route traffic of priority level 0 with an average cell loss probability of  $10^{-22}$ , traffic of priority level 1 with an average cell loss probability of  $10^{-18}$ , traffic of priority level 2 with an average cell loss probability of  $10^{-16}$ , and traffic of priority level 3 with an average cell loss probability of  $10^{-15}$ . Thus, it is quite apparent that the use of a high (and presumably more expensive) priority level can have a significant effect on the average cell loss probability that is experienced by a particular user.

The ATM switch 10A may be realized with either electronic or photonic/electronic components. Either embodiment may capitalize on the parallel processing architecture. An electronic embodiment is the more attractive presently because of the inexpensive and ease of procuring the components necessary construction. However, a photonic/electronic hybrid has great promise because of the possible advantages in connectivity. For a photonic/electronic embodiment, a novel interconnection technology known as free-space digital optics may be employed. System based on free-space digital optics use beams of light propagating roughly orthogonal to the plane of the device substrates to provide connections between devices. These beams are routed using lenses, mirrors, and holograms to the desired

receivers on the target device array. The free space optical approach to device connectivity may yield systems with very high degrees of connectivity between devices.

Although the ATM packet switch 10A was initially designed with the parallelism of free-space digital optics in mind, one of the advantages of the its architecture results from the fact that the switch fabric 14A can be implemented using any technology that can provide the functionality of many small, high-speed circuit switches. As a result, the architecture of switch fabric 14A is essentially technology-independent. An embodiment using free-space digital optics as the interconnection technology within the switch fabric is contemplated. The 16x16 crossbar switches within the switch fabric 14A will be implemented with FET-SEED device arrays. This approach may provide many benefits within the switch fabric 14A, because the resulting design based on optical interconnections may have lower levels of signal crosstalk, lower chip counts (and potentially lower cost) due to increased device integration, lower signal skew, and lower overall power dissipation (resulting in simpler thermal management techniques within the switch fabric 14A). These benefits might also prove to be useful in a free-space optical implementation of the out-of-band controller 30, although the design of the 640 Gigabits per second controller 30 is contemplated to be based on electronic technologies.

While the invention has been particularly illustrated and described with reference to preferred embodiments thereof, it will be understood by those skilled in the art that various changes in form, details, and applications may be made therin. For example, four or more levels of priority may be implemented instead of the two shown and described. It is accordingly intended that the appended claims shall cover all such changes in form, details and applications which do not depart from the true spirit and scope of the invention.

What is claimed is:

1. A packet switch for switching a telecommunication packet from a plurality of input lines to a plurality of output lines, comprising:

a plurality of input interfaces, each having an input port connected to a respective input line of said plurality of input lines, and each of said input interfaces having an output port;

a network for switching a plurality of I input ports to a plurality of P output ports;

each of said plurality of input interface output ports is fanned out to a respective group of F of said I input ports of said network,

said network having a plurality of C pipes, where C is an integer of a value equal to P/I;

a plurality of output modules, said output modules together having a plurality of inputs, each of said output module inputs connected to respective output port of said plurality of P output ports, and together having a plurality of outputs, each of said output module outputs connected to a respective output line of said plurality of output lines;

each pipe of said C pipes having a path from each of the plurality of inputs lines that is connectable to a respective output line of the plurality of output lines; and an out of band controller having multiple priority levels for hunting a path through said packet switch for a telecommunication packet.

2. The packet switch as set forth in claim 1, wherein: each of said input interfaces has a store for storing a telecommunication packet; and

25

said out-of-band controller rolls a request for a path for the telecommunications packet which was unable to find an unblocked path through a first pipe to an input of a pipe controller of a second pipe and the telecommunications packet is stored in said input interface while the controller is hunting an unblocked path.

3. The packet switch as set forth in claim 1, wherein: each of said input interfaces has a store for storing a telecommunication packet; and

said out-of-band controller rolls a request for a path for the telecommunications packet which was unable to find an unblocked path through both a first pipe and a second pipe to an input of a pipe controller of a third pipe and the telecommunication packet is stored in said input interface while the controller is hunting an unblocked path.

4. The packet switch as set forth in claim 1, wherein: each of said input interfaces has a store for storing a telecommunication packet; and

said out-of-band controller rolls a request for a path for the telecommunications packet which was unable to find an unblocked path through a first pipe, a second pipe and a third pipe to an input of a pipe controller of a fourth pipe and the telecommunication packet is stored in said input interface while the controller is hunting an unblocked path.

5. The packet switch as set forth in claim 1, wherein said out-of-band controller assigns an order of preference to ATM cells to reduce a probability of losing an ATM cell by internal blocking.

6. The packet switch as set forth in claim 1, wherein said out-of-band controller assigns an order of preference to ATM cells, and rolls a request for a path for a telecommunications packet which was unable to find an unblocked path through a first pipe to an input of a pipe controller of a second pipe.

7. The packet switch as set forth in claim 1, wherein said out-of-band controller assigns an order of preference to ATM cells, and rolls a request for a path for a telecommunications packet which was unable to find an unblocked path through a first pipe and a second pipe to an input of a pipe controller of a third pipe.

8. The packet switch as set forth in claim 1, wherein said out-of-band controller assigns an order of preference to ATM cells, and rolls a request for a path for a telecommunications packet which was unable to find an unblocked path through a first pipe, a second pipe and a third pipe to an input of a pipe controller of a fourth pipe.

9. A packet switch for switching telecommunication packets, comprising:

a network for switching a plurality of I inputs to a plurality of P outputs;

said network having a plurality of C pipes, where C is an integer of a value equal to P/I;

each pipe having a respective pattern of switching its inputs to its outputs;

each pattern of switching of a pipe is independent of the patterns of switching of the other pipes;

a plurality of output modules connected to said plurality of P outputs; and

an out of band controller having multiple priority levels for hunting a path through said packet switch for a telecommunication packet.

10. The packet switch as set forth in claim 9, wherein said out-of-band controller rolls a request for a path of a tele-

26

communication packet which was unable to find an unblocked path through a first pipe to an input of a second pipe.

11. The packet switch as set forth in claim 9, wherein said out-of-band controller rolls a request for a path of a telecommunication packet which was unable to find an unblocked path through both a first pipe and a second pipe to an input of a pipe controller of a third pipe.

12. The packet switch as set forth in claim 9, wherein said out-of-band controller rolls a request for a path of a telecommunication packet which was unable to find an unblocked path through a first pipe, a second pipe and a third pipe to an input of a pipe controller of a fourth pipe.

13. The packet switch as set forth in claim 9, wherein said out-of-band controller assigns an order of preference to ATM cells to reduce a probability of losing an ATM cell by internal blocking.

14. The packet switch as set forth in claim 9, wherein said out-of-band controller assigns an order of preference to ATM cells, and rolls a request for a path of a telecommunications packet which was unable to find an unblocked path through a first pipe to an input of a pipe controller of a second pipe.

15. The packet switch as set forth in claim 9, wherein said out-of-band controller assigns an order of preference to ATM cells, and rolls a request for a path of a telecommunications packet which was unable to find an unblocked path through a first pipe and a second pipe to an input of a pipe controller of a third pipe.

16. The packet switch as set forth in claim 9, wherein said out-of-band controller assigns an order of preference to ATM cells, and rolls a request for a path of a telecommunications packet which was unable to find an unblocked path through a first pipe, a second pipe and a third pipe to an input of a fourth pipe.

17. The packet switch as set forth in claim 9, wherein said out-of-band controller controls finds and establishes paths through said packet switch for the telecommunication packets.

18. An ATM switch for switching ATM packets, comprising:

a plurality of ATM interface cards, each having a respective input connected to an ATM telecommunication line and an output;

a network for switching a plurality of I inputs to a plurality of P outputs;

said network having a plurality of F pipes, where F is an integer of a value equal to P/I;

each of said outputs of said input interfaces is fanned out by a factor of F and connected a respective input of each of said F pipes;

each pipe having a respective pattern of switching its inputs to its outputs;

each pattern of switching of a pipe is independent of the patterns of switching of the other pipes;

a plurality of output modules connected to said plurality of P outputs;

each of said output modules having a plurality of outputs; and

an out of band controller having multiple priority levels for hunting a path from an input of one of said plurality of input interfaces through said ATM switch for an ATM packet to a desired output of one of said plurality of output modules.

19. The ATM switch according to claim 18, wherein an input interface of said plurality of input interfaces has a

memory that stores an ATM packet for two ATM packet periods to allow an ATM packet which was blocked during the previous ATM packet period to hunt a path from an input of one of said plurality of input interfaces through said ATM switch for an ATM packet to a desired output of one of said plurality of output modules during the next ATM packet period.

20. The ATM switch according to claim 19, wherein said ATM packet which was previously blocked for one ATM period is communicated through said ATM switch to its

respective desired output of its output module in proper sequence without the use of a time stamp.

21. The ATM switch according to claim 18, wherein an input interface of said plurality of input interfaces has a means for reading a priority field of each incoming ATM packet and means for releasing a request vector during a pipe hunting period assigned to its ATM packets according to the value in said priority field.

\* \* \* \* \*